Magyar Tudomány 179(2018)7, 964–976
DOI: 10.1556/2065.179.2018.7.5

A PLETYKA A TÁRSAS REND SZOLGÁLATÁBAN – AZ INFORMÁLIS KOMMUNIKÁCIÓ STRUKTÚRÁJÁNAK MÉLYEBB MEGÉRTÉSÉÉRT A *COMPUTATIONAL SOCIAL SCIENCE* ESZKÖZEIVEL[1]

GOSSIP IN SERVICE FOR SOCIAL ORDER – USING THE TOOLS OF COMPUTATIONAL SOCIAL SCIENCE FOR A DEEPER UNDERSTANDING OF THE STRUCTURE OF INFORMAL COMMUNICATION

Galántai Júlia[1], Pápay Boróka[2], Kubik Bálint György[3], Szabó Martina Katalin[4], Takács Károly[5]

[1] tudományos segédmunkatárs, MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoport
[2] tudományos segédmunkatárs, MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoport
[3] kutató, MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoport
[4] tudományos segédmunkatárs, MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoport, egyetemi tanársegéd, Szegedi Tudományegyetem Bölcsészettudományi Kar Orosz Filológiai Tanszék
[5] PhD, az MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoport vezetője
takacs.karoly@tk.mta.hu

ÖSSZEFOGLALÓ

A pletykálás, amelynek során másokról értékelő módon a hátuk mögött beszélünk, az emberi kommunikáció jelentős hányadát teszi ki. Miközben a mindennapi életben pejoratív értelemben beszélünk róla, a társadalomtudományi kutatások arra világítanak rá, hogy a pletykának jelentős szerepe van a társas rend biztosításában, a normák fenntartásában és a kooperáció biztosításában a különböző társas kontextusokban. Miközben ezeket a társadalomtudományi alapkérdéseket eddig főleg kvalitatív, önkitöltős kérdőíves vagy absztrakt kísérleti körülmények között vizsgálták, cikkünkben azt kívánjuk bemutatni, hogy miként gazdagíthatóak ezek a hagyományos elemzési módszerek a *computational social science* eszköztárának felhasználásával. Leírjuk azt a kvantitatív stratégiát, amellyel elemezzük a pletykát és alaptulajdonságait egy élő, spontán beszédet tartalmazó korpuszban. Bemutatjuk azokat a morfológiai, téma szerinti, szóhasználat szerinti, kontextuális és nem szövegalapú (például időzítés, beszélőváltás, nevetés és más hangok) elemzési lehetőségeket, amelyek a pletyka természetének és funkcióinak mélyebb társadalomtudományi megértését segítik.

[1] A kutatást az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében (ERC_CoG_2014_648693 sz. szerződésben) támogatja, a kutatás vezetője Takács Károly.

ABSTRACT

Gossip, as an evaluative talk about others who are not present, constitutes a major portion of human communication. Meanwhile we talk about gossip in a pejorative way in everyday language, research in social sciences highlights that gossip might play an important role in the establishment of social order, and in the maintenance of social norms and cooperation in various social contexts. These fundamental questions have been analyzed in the social sciences almost exclusively by qualitative methods, self-reports, and in abstract laboratory experiments. In this paper, we would like to illustrate, how these traditional research methods could be supplemented with the application of the tools of computational social science. We describe the quantitative strategy for the analysis of gossip and its characteristics in a large corpus of spontaneous conversations. We portray the morphological, topic-based, vocabulary-based, contextual, and non-textual (e.g., timing, turn-taking, laughing, and other sounds) opportunities of analysis that could improve the scientific understanding about the nature and functions of gossip.

Kulcsszavak: pletyka, társas normák, spontán beszéd, szövegkorpusz, korpuszépítés, kvantitatív szövegelemzés

Keywords: gossip, social norms, spontaneous speech, text corpus, corpus building, quantitative text analysis

BEVEZETÉS

Az emberi informális kommunikáció jelentős részét, egyesek szerint kétharmadát, jelen nem lévő, más személyekről folytatott értékelő tartalmú beszélgetés teszi ki (Dunbar, 1996, 2004; Foster, 2004). Az ilyen beszélgetéseket, amelyekben legalább egy értékelő és egy hallgató vesz részt, tekintjük pletykának (Kurland–Pelled, 2000; Ellwardt, 2011). Jó pletykát hallani és pletykálni mindenki szeret, mégis magához a pletykához a köznapi értelemben pejoratív konnotációkat fűzünk, elítéljük azt. Miért létezik akkor, és miért olyan elterjedt a pletyka? Miért használunk ki szinte minden alkalmat mások hátuk mögött történő kibeszélésére?

Ezeknek a kérdéseknek a megválaszolásához elengedhetetlen elsőként annak igazolása, hogy tényleg ilyen léptéket ölt-e az emberi társas kommunikációban a pletyka. Az erre vonatkozó ismeretanyag viszonylag szerény, és kvalitatív tapasztalatokra hagyatkozik. Nagyon ritka a spontán informális beszélgetéseket tartalmazó korpusz, amely ennek becslésére objektív lehetőséget adna. A magyar nyelvre vonatkozóan ilyen korpusz korábban nem létesült.

A pletyka gyakoriságának leírásán túlmutatóan szükség van arra, hogy megértsük, miért pletykálunk másokról, és miért van szükség ilyen mértékben mások értékelésére. Kutatási projektünk kiinduló hipotézise szerint azért, mert a pletykának pozitív közösségi funkciója van. A pletyka egy olcsó eszköz, amely

biztosítja a közösség szereplőinek a reputációs kontrollját, értesít azok esetleges normaszegéseiről, és így hozzájárul a társas normák fenntartásához, a közösségi rendhez, és elősegíti az együttműködést.

Mindezeket a tartalmakat a beszélt nyelvben a pletyka beazonosítása után lehet vizsgálni. Változtatja-e a hallgató a pletyka tárgyáról alkotott véleményét? Például: tesz-e olyan előjelű értékeléseket az adott személyről, amelyek egybevágnak a pletyka előjelével? Tartalmaz-e a pletyka utalást konkrét normaszegésre, kapcsolja-e a hallgató a pletykát az uralkodó közösségi normákhoz? Óvatosabban viszonyul-e a későbbiekben a hallgató a negatív pletyka tárgyát jelentő személyhez? Kerüli-e esetleg az együttműködést vele?

Ezen kérdések megválaszolását azonban számos lépés előzi meg, tele elméleti és gyakorlati buktatókkal. Már a pletyka azonosítása sem egyszerű feladat, a legtöbb felvetett kérdés vizsgálatához pedig az adott közösség társas viszonyainak hosszabb távú és emellett mély megfigyelésére van szükség. A jelen tanulmányban bemutatjuk, hogy miként lehet egy mélyebb megfigyelésből származó, nagy mennyiségű, élő spontán beszédet tartalmazó szövegkorpuszt a computational social science eszköztárával felhasználni ezeknek a kérdéseknek a vizsgálatához. Elsőként érvelünk amellett, hogy a mindennapi nyelv elemzésének és a korpuszok használatának óriási jelentősége lehet a társadalomtudományban alapvető kérdések vizsgálatakor. Ezt követően bemutatjuk a HuTongue korpusz létrehozásának legfontosabb lépéseit. Felvázoljuk a készülő adatbázis tervezett szerkezetét, a főbb változókat, majd az elemzési irányokat. Végül a továbblépési lehetőségek között tárgyaljuk azokat a további társadalomtudományi kérdéseket, amelyek vizsgálata megvalósulhat a HuTongue korpusz létrejöttével.

A MINDENNAPI NYELV ÉS A KORPUSZOK JELENTŐSÉGE A TÁRSADALOMTUDOMÁNYBAN

A különböző, így többek között a társadalomtudományi és az alkalmazott nyelvészeti (például pragmatikai) tárgyú kutatások egyik legfontosabb vizsgálati eszközét a számítástechnikai eszközökkel elemezhető formátumú szövegkorpuszok jelentik. A korpuszok három legfontosabb tulajdonságát a következőkben határozhatnánk meg: mindenekelőtt a korpusz ténylegesen előforduló írott vagy leiratozott beszélt nyelvi adatok gyűjteménye (Oravecz et al., 2014). Mindenképpen olyan adatokból áll tehát, amelyek a nyelvhasználat folyamán keletkeztek, nem pedig a kutató maga állította elő azokat a vizsgálat céljából, ún. introspektív adatként (vö. McEnery, 2005). A korpusz e sajátsága kétségkívül új távlatokat nyit a tudományos vizsgálatokban, hiszen annak tartalma – szemben például a nyelvészeti kutatásban egyébként gyakran alkalmazott int-

rospektív adatokkal – a nyelvi valóságot, a tényleges emberi nyelvhasználatot tükrözi, annak kvalitatív és kvantitatív sajátságaival együtt. Ha a problémát a társadalomtudományi kutatás oldaláról nézzük, még nyilvánvalóbbá válik a korpuszadat jelentősége: szociológiai vizsgálatokat ugyanis kizárólag valós, a természetes kommunikáció folyamatában létrejött nyelvi produktumokon lehet végezni. A szövegkorpuszok másik fontos előnye, hogy nagy mennyiségű szöveget foglalnak magukban, ezáltal általános érvényű állítások empirikus igazolását teszik lehetővé. Végül, az utóbbi sajátsággal összefüggésben, a szövegkorpuszok elektronikus formájúak, amelynek köszönhetően a bennük lévő nagy mennyiségű adat elemzése nem csupán manuális, de automatizált módszerekkel is lehetséges, így egy-egy sajátság feldolgozása gépi megoldással gyorsan és költséghatékonyan elvégezhető.

A Magyar Nemzeti Szövegtár (MNSZ2, URL1; Oravecz et al., 2014) honlapja alapján a korpuszban „a szövegeket valamilyen szempont szerint válogatják és rendezik. Nem feltétlenül egész szövegeket tartalmaz, és nemcsak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket (bekezdés, mondat).” Bár a definíció nem mutat rá, a szövegkorpuszok a legtöbb esetben valamilyen manuális vagy automatikus feldolgozási folyamaton (másképpen: annotáción) esnek át. Ráadásul ennek a feldolgozási folyamatnak a sajátságait a korpusz jövőbeli felhasználási céljai határozzák meg. A korpuszok annotációja egy rendkívül értékes sajátság, ugyanis ez teszi lehetővé a korpusz sok szempontú, géppel támogatott lekérdezését. Az annotáció olyan annotálási jelek (számítógépes nyelvészeti terminussal: tagek) alkalmazását jelenti, amelyeket a korpuszban levő szövegekre visznek fel a korpusz építői (másképpen annotátorai). Ezek a jelek hivatottak explicitté tenni a nyelvi adatokban már meglevő, azonban addig implicit formájú információt (vö. McEnery, 2005). Amennyiben a korpusz egyáltalán vagy az aktuális tudományos kutatásnak megfelelő annotációt nem tartalmaz, úgy lekérdezése, azaz a benne szereplő adatok automatikus kinyerése rendkívüli mértékben korlátozott, jobbára csupán kézi megoldással lehetséges.

A korpuszok annotálása automatikus, félautomatikus, valamint manuális munkával valósítható meg (vö. McEnery, 2005). Számos feladat (például a szótövezés vagy a szófaji egyértelműsítés) ma már olyan hatékonysággal végezhető el automatikus módszerrel, hogy ezekben a feladatokban nem szükséges humán annotátorokat alkalmazni. Amennyiben azonban az annotálás – annak jellege miatt – automatikus módszerrel egyáltalán nem végezhető el, és nem támogatható, úgy a teljes munkát humán annotátoroknak kell elvégezniük. Ezt a megoldást alkalmazzák a legtöbbször olyan esetekben, amikor a korpuszt valamely szemantikai vagy pragmatikai jelenség vizsgálatára kívánják felhasználni a jövőben. Ilyenkor ugyanis olyan nyelvi sajátságokat kell annotálni a korpusz szövegeiben, amelynek az automatikus feldolgozására jelenleg nincsenek ren-

delkezésre álló eszközeink. Tekintettel arra, hogy ezeknek a korpuszoknak az elkészítése jelentős költséget igényel, a nagyméretű, kézzel annotált adatbázisok száma kimondottan csekély.

A fentebb elmondottakkal összefüggésben a manuális módszerrel feldolgozott és sokrétű annotációval ellátott korpuszok két alapvető okból is a legértékesebb kutatási és fejlesztési eszközök között tartandók számon. Egyrészt tudományos felhasználásuk új, eleddig ismeretlen összefüggések feltárását teszi lehetővé. Másrészt a felhasználásukkal lehetségessé válhat újabb típusú annotációra is képes automatikus nyelvfeldolgozó eszközök fejlesztése és tesztelése.

A korpuszok között írott és beszélt nyelvi szövegkorpuszokat is találunk, azonban a legtöbb létező korpusz az írott nyelvet reprezentálja (McEnery, 2005). Ennek talán a legfontosabb oka az, hogy a beszélt nyelvi anyag feldolgozására jelenleg sokkal kevesebb eszköz áll a rendelkezésünkre, mint az írott nyelvi adatok kezelésére. Különösen csekély a magyar beszélt nyelvet reprezentáló korpuszok száma, és ezek is többségükben olvasott szövegeket tartalmaznak (vö. Gósy et al., 2012). Spontán, beszélt magyar nyelvet tartalmaz az adatközlői történetelmeséléseket tartalmazó Kivi korpusz (Kugler, 2015), az ún. Magyar spontán beszéd adatbázis (BEA) (Gósy et al., 2012), és egy készülőben lévő spontán beszédet tartalmazó korpusz, a Budapesti Egyetemi Kollégiumi Korpusz (BEKK, URL2). A BEA korpusz létrehozóinak fő célja az volt, hogy fonetikai, és nem szemantikai vagy pragmatikai vizsgálatokat tegyen lehetővé, ennek megfelelően alakították ki a korpuszban alkalmazott annotációt. Az elmondottak okán a BEA-korpusz társadalomtudományi tárgyú kutatásokra csupán korlátozottan alkalmazható. A Bodó Csanád kutatócsoportja által épített BEKK Budapesten élő fiatalok egymás közötti nyelvi interakcióit tartalmazza, és eleve társadalomtudományi céllal készül. A társalgáselemzés keretében lehetővé teszi majd a társas identitások létrejöttének és változásának vizsgálatát, különösen a társadalmi nem és a szexualitás konstrukciói mentén.

Míg a BEKK esetében az interakciókat a résztvevők saját telefonjaikon rögzítették, ezért szelektív társalgásokat tartalmaz, és nem reprezentálja tökéletesen a teljes élőbeszédet, az általunk létrehozandó HuTongue korpusz hiánytalanul tartalmaz majd egy hosszabb időszakból magyar nyelvű spontán beszélgetéseket. A nagyméretű és megfelelően annotált korpusz lehetővé teszi a pletyka eleddig ismeretlen természetének a feltárását, mélyebb megismerését. Emellett az adatbázis, reményeink szerint, különböző célú további szociológiai és nyelvészeti tárgyú kutatáshoz és fejlesztéshez nyújt majd táptalajt a jövőben.

A HUTONGUE KORPUSZ LÉTREHOZÁSA

A hanganyag előfeldolgozása

A HuTongue korpusz alapja egy hozzávetőlegesen ezerórás hanganyag, amelyet a nap 24 órájában rögzítettek nyolc napon keresztül, zárt környezetben.[2] A nyolc résztvevő mindegyike saját mikrofonnal rendelkezett, ezzel segítve elő a hangfájlok jó minőségű rögzítését. A hanganyagot kiváló minőségben tartalmazó tömörítetlen audiofájlok összmérete mind a hanganyag exportálását, tárolását, mind a feldolgozását megnehezítette, így eleve kisebb méretre hozott .wav kiterjesztésű audiofájlokkal dolgoztunk. Az audiofájlok tárolására és feldolgozására az MTA Felhőben létrehozott virtuális gépen kaptunk lehetőséget.

A hanganyag a résztvevők ébren töltött idejének és a hosszú csendek, nem beszédhangok (mosogatás, háttérzene stb.) szűrése után összesen közel ötszáz órányi leiratozható hangfelvételt tartalmaz. Az előfeldolgozás során először eltávolítottuk azokat a részeket (néhány esetben egész órákat), melyek közel teljes csendet tartalmaztak, egy hangerősségbeli küszöbérték felhasználásával. A szűrés következő lépcsője során egy gépi tanulást alkalmazó módszerrel azonosítottuk azokat az egységeket, amelyek nem minősülnek csendnek, avagy kellően hangosak voltak, és koherens szegmenseket képeztek. Mivel az azonosítás sok esetben kis terjedelmű (néhány tizedmásodperces) szegmenseket eredményezett, ezeket a bennük hallható emberi hang (a Voice Activity Recognition technika alkalmazásával) és időbeli eloszlásuk alapján (egymáshoz képest 15 másodperces távolságban) nagyobb méretű szegmensekké egyesítettük. A nagyobb egységeket aztán darabonként körülbelül egyórás audiofájlokban fűztük össze, melyekben a szegmenshatárokat jól elkülöníthetően egy dallammal jeleztük. A szűrés egyes lépéseit manuálisan ellenőriztük, hogy az előfeldolgozás eredménye koherens egységeket alkosson, és hogy kivehető emberi beszédet a szűrőmechanizmus semmilyen esetben se távolítson el. Az ellenőrzés a módszer magas pontosságára engedett következtetni.

A gépi leiratozás próbái sajnos nem hoztak megbízható eredményeket, így a HuTongue magyar nyelvű, spontán nyelvi korpusz kézi leiratozással és annotálással történő építése mellett döntöttünk.

[2] A spontán nyelvi hanganyagot, amelyet egy szórakoztatóipari cég rögzített, kizárólag tudományos célokra adták át, és használjuk fel, teljes titoktartási kötelezettségvállalás mellett. A hanganyag a résztvevők nyolc napra vonatkozó összes beszélgetését tartalmazza. A felvételen részt vevő önkéntesek teljes körű tájékoztatásban részesültek a hangfelvételek elkészüléséről.

Annotáció

A hanganyagot a különböző időszegmensek visszakereshetővé tétele és összefűzése érdekében meghatározott tagolás után az annotálók időbélyeggel látták el. Az így kapott nyersanyag azonkívül, hogy megőrzi a hanganyaggal való kapcsolatát, további időintervallum alapú mérési eszközök kidolgozására ad lehetőséget. A kézi annotálás lehetőséget biztosít ahhoz, hogy kiszűrjük a szövegből azokat az értékelő párbeszédeket, amelyek során a diskurzus tárgya nincs jelen, tehát az adott beszélgetést pletykaként értékelhetjük. A pletyka megjelenésekor azonosíthatóvá válik a küldő és fogadó fél, hiszen az annotátorok kódokkal jelzik a résztvevők nevét és a pletyka tárgyát is, amennyiben a résztvevők közül kerül valaki említésre. A diskurzus közben beazonosíthatunk olyan további személyeket is, akik a beszélgetés során csendben maradtak, de jelenlétük az annotáló által érzékelhető.

A korpusz kialakításakor arra törekedtünk, hogy mélyebb betekintést nyerjünk a pletyka spontán beszédhelyzetekben való megnyilvánulásairól, ezért annotációs jeleket használtunk a beszélők érzelmi megnyilvánulásának rögzítéséhez. A kézi annotálás nem szövegszerű kódok rögzítésére is lehetőséget ad. Az így használt annotációs kódok egyik csoportja az emóciós hanghatások lejegyzéséhez fűződik, hiszen arra is kíváncsiak vagyunk, hogy melyek azok az érzelmek, amelyek a mindennapi, informális kommunikáció során megjelenhetnek. Az ilyen annotációs jelek például a nevetés, sírás, torokköszörülés, sóhajtás, gúnyos nevetés stb. azonosítására alkalmasak (Szabó–Galántai, 2017). Így vizsgálhatóvá válhat, hogy milyen típusú érzelmek jelennek meg olyan mindennapi, társas szituációkban, amelyekben pletyka hangzik el.

További annotációs jelek beazonosíthatóvá teszik a beszélőváltást, a beszéd közbeni szünetek hosszát és gyakoriságát, az egy időben történő beszédet és az egymás szavába vágás előfordulásának jellemzőit is, amelyek a beszélők közötti erőviszonyok feltárását is lehetővé teszik.

A korpusz minőségének és értelmezhetőségének biztosítása érdekében értelmező annotációs kódokat is alkalmazunk, melyek az érthetetlen, az azonosíthatatlan vagy nem a diskurzusban részt vevő személyektől érkező beszédet jelölik (részletesebben: Szabó–Galántai, 2017).

Minőségbiztosítás

Az így kapott korpusz megbízhatóságát és összeegyeztethetőségét többféle módszerrel és több dimenzióban mérjük. Mindez a kvalitatív, szúrópróbaszerű ellenőrzés mellett gépi eszközökkel történik. A leiratozást és annotálást végzők munkáját több dimenzióra bontva hasonlítjuk össze, egyrészt egymáshoz képest, másrészt egy referenciagépelőhöz viszonyítva. Súlyos minőségi kifogások ese-

tén az adott szöveganyagot újra leiratoztatjuk és annotáltatjuk. A leiratozóknak és annotálóknak az egyes minőségbiztosítási dimenziókban egyéni visszajelzést adunk, és munkájuk minőségjavulását ellenőrizzük. Az összehasonlítás a szövegegyezés, az annotálás, a szereplők azonosítása és jelölése, és az időbélyeg használatának fő dimenzióiban történik, amelyeket a pontosabb visszajelzés érdekében részdimenziókra bontunk. Ezeket a korpuszépítés alatt folyamatosan ellenőrizzük, és mérőszámokkal dokumentáljuk. A szövegegyezés főbb mérőszámaiként a tisztított szegmenseken számolt Levenshtein-távolságot (Levenshtein, 1965) és a koszinusz hasonlóságot (cosine similarity) használjuk.

Az adatbázis szerkezete

A nagy szövegkorpusz eltárolása egy olyan adatbázisban történik, amely hatékonyan képes nagy mennyiségű szöveget rögzíteni és kereshetővé tenni (Elasticsearch). A szöveges keresőmotor képes a teljes korpuszunk megbízható tárolására és gyorskereső, összegző és akár elemzési műveletek végrehajtására is.

Az általunk gyűjtött információmennyiség több különböző (az Elasticsearch terminológiáját használva) index és típus alatt tárolódik az adott adattípus jellemzőitől függően. Két fő adatsort különböztetünk meg. Az első adatbázist az egyedi (egy szereplőtől egy időegységben származó) megszólalások időbélyegekkel (a megszólalás pontos ideje az adatfelvétel időintervallumán belül) ellátott korpusza adja. Az egyes megszólalásokhoz az időbélyegeken túl számos egyéb attribútumot társítunk, melyek a beszélő azonosítóját, az egyes annotációs kódok jelenlétét, a szöveget rögzítő gépelő kilétét, és még sok más információt tartalmaznak. A teljes szövegkorpusz adatbázisát kiegészítendő létrehozunk egy olyan adatsort is, amely a különálló szövegrészek minden egyes szavához olyan attribútumokat társít, mint a szófaja, a mondatrész vagy a szó lemmatizált formája. Ezen adatbázis minden eleme egyértelműen megfeleltethető a korpusz minden szavának. Ez a struktúra a kutatók számára jelentősen megkönnyíti a kereshetőséget és az elemzést, különösen az NLP (Natural Language Processing) problémákra vonatkozóan.

Az adatbázis továbbá lehetővé teszi, hogy a beszélők hangfájljaiból leiratozott beszédeket párbeszéddé illesszük össze. Ezzel a párbeszéd is elemzési egységgé válhat a későbbiekben. Mivel minden beszélőnek saját hangfelvevő készüléke volt, ezért sok esetben ezek leiratai ugyanazon párbeszéd különböző részeit tartalmazzák. Mivel minden leiratozott sor saját időbélyeggel rendelkezik, az azonos időben történő beszédeket egymás mellé tesszük, és az illesztést horgonyszavak segítségével pontosítjuk. Ezt segíti, hogy minden leirat tartalmaz több beszélőt, és tartalmazza, hogy kik vettek részt az adott diskurzusban. A pletyka azonosításának szempontjából zárt közösségről lévén szó, ugyancsak fontos, hogy ennek következtében az is behatárolható, hogy kik *nem* voltak jelen az adott beszélgetés során.

A PLETYKA ELEMZÉSÉNEK A COMPUTATIONAL SOCIAL SCIENCE ÁLTAL KÍNÁLT IRÁNYAI

A korpusz tervezett automatikus feldolgozási lépései

A manuálisan gépelt és annotált korpuszt különböző automatikus megoldásokkal is fel kívánjuk dolgozni. Ahogyan azt a korpusz feldolgozása kapcsán ismertettük (lásd fentebb), a létrejövő spontán nyelvi adatbázis tartalmaz szövegszerű és nem szövegszerű változókat is. Az automatikus feldolgozási megoldásokkal közülük a szövegszerű adatokra szeretnénk további információkat felvinni az elkészült adatbázisban. E munka célja, hogy a korpusz minél részletesebb annotációval rendelkezzen a szövegek grammatikai, szemantikai, valamint pragmatikai sajátságairól (Szabó–Galántai, 2017).

A szövegszavak tokenizálásához, morfológiai, valamint szófaji elemzéséhez a magyarlanc elemző eszközt kívánjuk használni (Zsibrita et al., 2013). Ez a feldolgozási lépés két okból is fontos a számunkra. Egyrészt úgy véljük, hogy a különböző grammatikai sajátságok kvantitatív adatai a pletykaszövegek azonosításában is hasznosíthatóak lesznek (lásd lentebb). Másrészt, a magyarlanccal kapott kimenet megfelelő bemenetként szolgálhat a további automatikus szemantikai feldolgozási lépésekhez.

A szöveg szemantikai tartalmát illetően elsősorban névelem-felismerést, szentiment- és emócióelemzést, valamint topikmodellezést tervezünk végrehajtani a kész korpuszon. A szentiment- és emócióelemzéshez a természetesnyelv-feldolgozás (NLP) eszköztárából a szótárillesztés módszert alkalmazzuk, amely például a gépi tanulás, a szintaktikai elemzésen alapuló mintaillesztés mellett egyszerűbb és költséghatékonyabb információkinyerési módszer (Drávucz–Szabó, 2017). A szentimentelemzéshez olyan lexikonra van szükségünk, amely szótári formába rendezi a lexikai szinten pozitív vagy negatív értékelő tartalommal rendelkező nyelvi elemeket. A szövegszintű emóciók felcímkézéséhez pedig olyan szótárat kell alkalmaznunk, amely a különböző érzelmek nyelvi realizációit tartalmazza, illetve minden olyan elemet, amely valamely emóció meglétére utalhat (Szabó–Morvay, 2015). E két elemzés segítségével fel tudjuk tárni a szövegekben megbúvó negatív és pozitív értékítéleteket és a különböző emóciókat.[3]

A korpuszban megjelenő témákat két megoldással kívánjuk feltárni. Annak érdekében, hogy a diskurzusok témáit leíró módon meghatározhassuk, és ezt követően tovább elemezhessük, az elkészült korpusz segítségével saját nyelvi szótárakat is készítünk. A szótárak a tervezetteknek megfelelően egy- és többszavas kifejezéseket egyaránt tartalmaznak majd. A szótárakon kívül topikmodellek segítségével tervezzük a nagyméretű szöveghalmazban a különböző témaegységek elkülönítését (például: pletyka, politika, sport, időjárás stb.).

[3] Erre a célra például a Precognox Informatikai Kft. által fejlesztett szótárak használhatóak.

Ahhoz, hogy a korpusz szövegeinek pragmatikai sajátságait még mélyrehatóbban feltárhassuk, a korpuszt a nyelvi bizonytalanság különböző típusú jelölőinek a szótárával, valamint a diskurzusjelölők szótárával is fel kívánjuk dolgozni. A bizonytalanságot jelölő kifejezések automatikus azonosítása napjaink nyelvtechnológiai kutatásainak egyik fontos problémaköre (Vincze, 2014). A bizonytalanságot jelölő elemek ugyanakkor a beszélői szubjektivitás fontos indikátorai lehetnek, ezért a nyelvi bizonytalanság detektálásával lehetőségünk nyílhat a vizsgált szemantikai tartalmak szubjektivitási értékeinek a megismerésére is (Drávucz–Szabó, 2017). A feldolgozás során egy bizonytalanságjelölő elemeket tartalmazó szótárat (Vincze, 2014) illesztünk a kész korpuszra. Mindez a pletyka értékítéletet tartalmazó tulajdonságát hivatott azonosítani.

Végezetül, a diskurzusjelölők annotálása lehetőséget teremt arra, hogy a diskurzusok e fontos szövegszintű kapcsolóelemeit is lekérdezzük, és azok kvantitatív és kvalitatív sajátságait a spontán nyelvhasználatban feltárjuk. A diskurzusjelölőknek nincs lexikális értelemben vett jelentésük, ehelyett procedurális (műveleti) jelentéssel bírnak. Színesítik, árnyalják a mondandót, illetve emocionális tartalmakról és beszédtervezési folyamatokról tanúskodnak.

A pletyka azonosítási lehetőségei

A fentebbieknek megfelelően, a manuálisan és gépi megoldásokkal egyaránt feldolgozott adatbázis számos elemzési módot kínál a kutatási kérdéseink megválaszolására. A pletyka általunk vázolt és az irodalomban általánosan elfogadott definíciója alapján az azonosításhoz több feltétel szükséges. Egyrészt, szükséges annak megállapítása, hogy a beszélő egy harmadik személyről beszél. Másrészt, szükséges annak megállapítása, hogy a harmadik személy nincs jelen a beszélgetésnél. Harmadrészt, szükséges annak meghatározása, hogy az említés értékelő tartalommal jár együtt. Az azonosítás lépései többféleképp történhetnek, ezért pragmatikusan megkülönböztethetünk pletykát szűkebb és tágabb értelemben is.

A legegyszerűbb általunk használt módszer az annotálás során használt jelekre és a névemlítés személyes és utaló névmásokkal kiegészített összekapcsolására épül. Az azonosítást követően a gazdagon annotált korpusz lehetővé teszi számunkra, hogy azonosítsuk a pletyka megjelenésében potenciálisan szerepet játszó tényezőket, és mérjük ezek fontosságát. Az elemzés nagyrészt a kvantitatív szövegelemzés eszközeivel történik.

A pletyka és nem pletyka jellegű szövegeket arányaiban és karakterisztikáiban is el tudjuk majd különíteni egymástól. A részletesen bemutatott manuális és automatikus feldolgozási megoldások alapján, prediktív modellek segítségével képesek leszünk megvizsgálni, hogy milyen tényezők valószínűsítik a pletykát egy adott szövegben. Ilyen a különböző nyelvi kifejezések jelenléte (például: a

szentiment- és emóciókifejezések, a nyelvi bizonytalanság jelölői vagy a diskurzusjelölők), a lexémák morfológiai és szófaji tulajdonságai, valamint a különböző, a szövegekben megjelenő nem szövegszerű hanghatások. Mindemellett a topikmodellezéssel feltárt témák sajátságai rámutathatnak, hogy a pletykához milyen témák kapcsolódnak, és melyek azok a témák, amelyek tipikusan nem jelennek meg a pletykaként azonosított szövegek közvetlen környezetében.

A topikmodellezés célja a beszélgetések tartalmában a közösségi normákra, együttműködésre utaló tartalmak azonosítása is. Ez hozzájárulhat annak a kiinduló hipotézisnek a teszteléséhez, hogy a pletyka pozitív szerepet játszik a közösségi normák és a kooperáció fenntartásában.

A vizsgált jelenség sajátságai okán elsősorban szövegcsoportosításra alkalmazott klasszifikációs modellekkel fogunk dolgozni. A szövegkategorizáció a nyelvi elemek előre meghatározott kategóriákhoz való hozzárendelése a tartalmuk figyelembevételével. A pletyka szempontjából lényeges tartalmak azonosítása egy csoportosítási problémának tekinthető, amely megvalósítható neurális hálózatok modelljeivel, *support vector machines* (SVM) vagy naiv bayesi klasszifikációs modellek segítségével.

Az elkészült adatbázisban lehetséges lesz a beszélők informális kommunikációs hálózatának kvantitatív elemzése. A kapcsolatok alakulását és a hálózat változását időben is vizsgálni tudjuk kapcsolati háló elemzési módszerekkel, valamint azonosítható és vizsgálható lesz konkrét tartalmú értékítéletek terjedése az adott kisközösségben.

A pletyka mint diskurzus

Az adatbázis lehetővé teszi, hogy az egyes diskurzusok között időbeni összekapcsolásokat végezzünk. Az összekapcsolás a gyakorlatban azonban egyáltalán nem triviális. Megnehezíti egyrészt az egy időben, de egymástól távol zajló beszélgetések azonos időbélyege és ugyanazon beszélgetés két oldalról történt egymástól eltérő leírása. Az összekapcsoláshoz az időbélyegek használata mellett ezért a beszélők és a beszélőpartnerek annotált jelzéseit és a szövegben előforduló horgonyszavakat és horgonyszóláncokat is használjuk.

Az összeillesztett párbeszédben lehetőség nyílik diskurzuselemzésre, ahol megvizsgálhatjuk, hogy milyen stratégiákat alkalmaz a beszélő, és hogyan vonja be a fogadó felet a pletykálás folyamatába. Azonosíthatókká válnak a fogadó fél tipikus reakciói és saját értékítéletének a változása.

Mivel a korpusz mellé a hangfájlok is rendelkezésre állnak, a későbbiekben a megfelelő elemzési módszerek fejlődésével és az annotált leirattal való összekapcsolással a hangok karakterisztikáit is megvizsgálhatjuk, mint például, hogy milyen sűrűn, milyen hangosan vagy milyen intenzitással beszélnek az azonosított pletykahelyzetekben a beszélők.

ÖSSZEGZÉS

Általánosabban arra szeretnénk választ kapni, hogy a mindennapi élet során hogyan jelenik meg a pletyka? Kézzelfogható eszközeink vannak arra, hogy olyan alapállapításokat is felülvizsgáljunk, amelyeket eddig a szakirodalomban készpénznek vettek, de megnyugtató módon empirikusan nem támasztottak alá, mint hogy az emberek a beszélgetéseik kétharmadát pletykálással töltik, vagy hogy a pletyka alapvető funkciója a normaszegők kibeszélése.

A pletyka operacionalizálásához kapcsolódó kérdésünk, hogy miként azonosítható be kvantitatív szövegelemzési eszközökkel az élő, spontán szövegben a pletyka, hogyan lehet azt elválasztani a diskurzus más elemeitől és témáitól, és milyen morfológiai jellemzőkkel írható le.

IRODALOM

Drávucz F. – Szabó M. K. (2017): A beszélői szubjektivitás vizsgálata szentiment- és emóciókorpuszokon. In: Ludányi Zs. (szerk.): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017. XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia.* Budapest, 39–49. http://www.nytud.hu/alknyelvdok17/proceedings/Dravucz_Szabo.pdf

Dunbar, R. I. (1996): *Grooming, Gossip and the Evolution of Language*. Cambridge, MA: Harvard University Press

Dunbar, R. I. (2004): Gossip in Evolutionary Perspective. *Review of General Psychology*, 8, 2, 100–110. DOI:10.1037/1089-2680.8.2.100, http://allegatifac.unipv.it/ziorufus/Dunbar%20gossip.pdf

Ellwardt, L. (2011): *Gossip in Organizations. A Social Network Study. (ICS Dissertation Series*) Groningen. https://www.researchgate.net/publication/254821799_Gossip_in_organizations_a_social_network_study

Foster, E. K. (2004): Research on Gossip: Taxonomy, Methods, and Future Directions. *Review of General Psychology.* 8, 2, 78. DOI:10.1037/1089-2680.8.2.78, https://pdfs.semanticscholar.org/8b2f/3c70bd2346b2218b743a765766e5a80a1718.pdf

Gósy M. – Gráczi T. E. – Gyarmathy D. et al. (2012): *Magyar spontán beszéd adatbázis = Hungarian Spontaneous Speech Corpus.* OTKA kutatási beszámoló, http://real.mtak.hu/12552

Kugler N. (2015): *Megfigyelés és következtetés a nyelvi tevékenységben.* Budapest: Tinta Kiadó

Kurland, N. B. – Pelled, L. H. (2000): Passing the Word: Toward a Model of Gossip and Power in the Workplace. *Academy of Management Review*. 25, 2, 428–438. DOI: 10.5465/AMR.2000.3312928, http://www.csun.edu/~nkurland/PDFs/AMR%20Gossip%202000.pdf

Levenshtein, V. I. (1965): Двоичные коды с исправлением выпадений*, вставок и замещений символов. Доклады Академий Наук СССР,* 163, 4, 845–848. http://www.mathnet.ru/links/50defca5677a80b2d88d3dc027ac4173/dan31411.pdf, Angolul: Levenshtein, V. I. (1966): Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady.* 10, 8, 707–710. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf

McEnery, T. (2005): Corpus Linguistics. In: Mitkov, R. (ed.): *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford, 448–463.

Oravecz Cs. – Váradi T. – Sass B. (2014): The Hungarian Gigaword Corpus. *Proceedings of LREC*, *Reykjavík, European Language Resources Association (ELRA)*, 1719–1723. http://real.mtak.hu/20143/1/681_Paper.pdf

Szabó M. K. – Galántai J. (2017): Egy magyar nyelvű spontán beszélt nyelvi korpusz (HuTongue) létrehozásának tapasztalatai. In: *XXVI. MANYE Kongresszus konferenciakötete.* Pécs

Szabó M. K. – Morvay G. (2015): Emócióelemzés magyar nyelvű szövegeken. In: Gecső T. – Sárdi Cs. (szerk.): *Nyelv, kultúra, társadalom*. Budapest: Tinta Kiadó, 286−292.

Vincze V. (2014): Uncertainty Detection in Hungarian Texts. In: *Proceedings of COLING 2014.* Dublin. 1844–1853. http://www.aclweb.org/anthology/C/C14/C14-1174.pdf

Zsibrita J. – Vincze V. – Farkas R. (2013): magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. *Proceedings of Recent Advances in Natural Language Processing*, Hissar. 763–771. https://www.aclweb.org/anthology/R/R13/R13-1099.pdf

URL1: http://clara.nytud.hu/mnsz2-dev/
URL2: bekk.elte.hu