StaffTalk: magyar nyelvű spontán beszélgetések korpusza

Szabó Martina Katalin[1,2], Vincze Veronika[3], Ring Orsolya[1], Üveges István[2,4], Vit Eszter[5,6], Samu Flóra[5,6], Gulyás Attila[1], Galántai Júlia[5], Szvetelszky Zsuzsanna[1], Bodor-Eranus Eliza Hajnalka[1], Takács Károly[5,1]

[1]Társadalomtudományi Kutatóközpont, CSS-RECENS
1097 Budapest, Tóth Kálmán utca 4.
[2]Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.
[4]Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
6722 Szeged, Egyetem utca 2.
[5]Linköpingi Egyetem, The Institute for Analytical Sociology
601 74 Norrköping, Svédország
[6]Budapesti Corvinus Egyetem
1093 Budapest, Fővám tér 8.
{martina,vinczev}@inf.u-szeged.hu
{Szabo.Martina,Ring.Orsolya,Gulyas.Attila,Bodor-Eranus.Eliza}@tk.hu
{uvegesistvan898,szvetelszky}@gmail.com
{eszter.vit,flora.samu,julia.galantai,karoly.takacs}@liu.se

Kivonat A cikkben bemutatjuk a StaffTalk nevű, nagy méretű, kézzel annotált korpuszt, mely magyar nyelvű spontán beszélgetéseket tartalmaz. A korpusz létrehozásával elsősorban ahhoz szerettünk volna vizsgálati anyagot teremteni, hogy zárt közösségeken belül az informális kommunikáció és a megbecsültség hogyan befolyásolja a közösség működését és normarendszerét. A munka első lépéseként a hanganyagokat legépeltettük, amelynek során a verbális információn túl egyéb, nem verbális információk megjelölésére is megkértük az annotátorokat. A legépelt hanganyagokat ezt követően három szinten annotáltuk: a beszélgetésekben megjelenő pletykát, beszédaktusokat és egyéb pragmatikai jegyeket, valamint bizonytalanságra utaló szavakat egyaránt megjelöltünk. Mindezeknek a sajátságoknak köszönhetően a kiinduló kutatási kérdéssel összefüggésben, valamint azon túl is a korpusz sokféle pragmatikai szempontú elemzés elvégzésére is alkalmassá vált.
Kulcsszavak: korpusz, spontán beszéd, kézi annotálás, pragmatika, szemantika, pletyka

1. Bevezetés

Manapság a társadalomtudományok területén is egyre népszerűbbé válnak a korpuszalapú, illetve számítógépes nyelvészeti eszközöket alkalmazó vizsgálatok. Je-

len kutatási programunk arra a kérdésre keresi a választ, hogy zárt közösségeken belül az informális kommunikáció és a megbecsültség hogyan befolyásolja a közösség működését és normarendszerét, és a kutatás során a pletyka mint diskurzus vizsgálatára fókuszál. A pletyka ugyanis jelentős szerepet játszik az interperszonális informális kommunikációban, korábbi kutatások szerint ezeknek legalább a felét, de akár kétharmadát is lefedheti (Dunbar, 1996, 2004; Foster, 2004).

A kutatás keretében okosórák segítségével hangfelvételeket készítettünk zárt közösségekben. Az itt bemutatott kutatási fázisban egy iskola oktatói karának diskurzusait rögzítettük, majd e felvételek leiratozását és annotálását végeztük el.

A tanulmányban részletesen bemutatjuk a hangfelvételek keletkezési körülményeit, az anyagok feldolgozási módszereit és eszközeit, valamint a StaffTalk korpusz alapvető statisztikai adatait.

2. Kapcsolódó irodalom

Az alábbiakban röviden áttekintjük a spontán beszélt nyelvi adatbázisokra, valamint a korpuszban alkalmazott annotációs szintekre vonatkozó szakirodalmat.

2.1. Beszélt nyelvi korpuszok

A különböző, így többek között a társadalomtudományi és a nyelvészeti (például pragmatikai) tárgyú kutatások egyik legfontosabb vizsgálati eszközét a számítástechnikai eszközökkel elemezhető formátumú szövegkorpuszok jelentik. A korpuszok között írott és beszélt nyelvi szövegkorpuszokat is találunk, azonban a legtöbb létező korpusz az írott nyelvet reprezentálja (McEnery, 2012). 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 (Galántai és mtsai, 2018).

Az elmúlt évtizedekben több spontán beszélt nyelvi korpusz keletkezett több nyelven is (Crowdy, 1993; Hemphill és mtsai, 1990; Maekawa és mtsai, 2000; Oostdijk, 2000; Van Bael és mtsai, 2007), amelyek között találunk agglutináló nyelveket reprezentáló korpuszokat is (Neuberger és mtsai, 2014), például török (Mengusoglu és Deroo, 2001) és finn (Seppänen és mtsai, 2003).

Ugyanakkor a korpuszok növekvő száma ellenére még mindig csak néhány van, amely hangzó szövegeket tartalmaz, azok gépelt leirataival együtt. Ez az átírási eljárás magas munkaerő- és költségigényével magyarázható. Különösen csekély a magyar beszélt nyelvet reprezentáló korpuszok száma, és ezek is többségükben felolvasott szövegeket tartalmaznak (Gósy, 2013). Az első magyar nyelvű beszélt korpusz a 20. század elején keletkezett (Neuberger és mtsai, 2014). Az elmúlt évtizedekben készült beszédadatbázisok többsége rögzített olvasott beszédet, vezetett történetmondást vagy irányított beszélgetéseket tartalmaz. A magyar telefonbeszéd-adatbázis (MTBA) egy beszédkorpusz, amely 500 alany által telefonon rögzített olvasott szövegből áll. Úgy tervezték, hogy támogassa a

beszédtechnológia területén végzett kutatásokat és fejlesztéseket (Vicsi és mtsai, 2002). Az úgynevezett Kivi korpusz (Kugler, 2015) különféle, videón látott történetek elmeséléseit tartalmazza, míg a Budapesti Szociolingvisztikai Interjú 250 adatközlő interjújából áll (Váradi, 2003). A HuComTech multimodális korpusz körülbelül 50 órányi video- és hangfelvételt tartalmaz 111 formális (szimulált állásinterjúk) és 111 informális, de irányított párbeszédből (Pápay és mtsai, 2011). A szövegek létrejöttének körülményei, valamint a szövegek feldolgozási módja miatt azonban a fentebbi korpuszok nem támogathatják a magyar spontán beszéd nyelvi sajátosságainak kutatását.

Legjobb tudomásunk szerint jelenleg három korpusz van, amely a magyar spontán beszédet kívánja reprezentálni, ez a Budapesti Egyetemi Kollégiumi Korpusz (BEKK) (Bodó és mtsai, 2017), a BEszélt nyelvi Adatbázis (BEA) (Gósy, 2013), valamint a HuTongue (Galántai és mtsai, 2018). Végezetül, meg kell még említenünk a CHILDES adatbázisból elérhető magyar gyereknyelvi korpuszokat is (Babarczy, 2009).

Ugyanakkor, a fenti korpuszok saját vizsgálataink elvégzésére nem voltak alkalmasak. Egyrészt, a BEKK esetében az interakciókat a résztvevők saját telefonjaikon rögzítették, ezért szelektív társalgásokat tartalmaz, ami nem reprezentálja tökéletesen az élőbeszédet. Másrészt, a BEA korpusz létrehozóinak fő célja az volt, hogy fonetikai, és nem szemantikai vagy pragmatikai vizsgálatokat tegyen lehetővé: így alakították ki a korpuszban alkalmazott annotációt. Az elmondottak okán a BEA-korpusz társadalomtudományi tárgyú kutatásokra csupán korlátozottan alkalmazható. Végül, az ún. HuTongue korpuszt (Galántai és mtsai, 2018) csupán félig (vagy részlegesen) spontánnak tekinthetjük, mivel egy szórakoztató jellegű tévéműsor céljaira készültek a felvételek, és, bár a társalgások a legtöbbször nem voltak kívülről kérdésekkel vagy témameghatározásokkal irányítva, a szövegek keletkezési körülményei (a résztvevők motivációi, valamint az időnkénti rendezői irányítás) befolyásolhatták a beszélői megnyilatkozásokat.

2.2. Pletykára annotált korpuszok

A pletykadiskurzusok elemzését, valamint társadalmi szerepét a korábbi kutatások többnyire kvalitatív vagy kérdőíves kvantitatív módszerekkel végezték (Eckhaus és Ben-Hador, 2019), írott szövegeken (Mitra és Gilbert, 2012) vagy megfigyelésen keresztül (Dunbar, 2004). A spontán előbeszéden belüli pletyka feltárására ezidáig kívül kevés korpusz és empirikus eredmény állt rendelkezésre (tudomásunk szerint kizárólag Robbins és Karan (2020)). Egyetlen magyar nyelvű korpuszról van tudomásunk (a fentebb említett HuTongue-ról), amely kifejezetten a pletyka természetének spontán beszéden belüli vizsgálatára irányul (Gulyás és mtsai, 2018; Szabó és Galántai, 2017), azonban a jelen tanulmányban bemutatott korpusz több szempontból jelentősen gazdagítja a pletyka vizsgálatának lehetőségeit a HuTongue korpuszhoz képest.

2.3. Pragmatikai annotációt tartalmazó korpuszok

Bár, amint azt korábban a 2.1. fejezetben említettük, a nemzetközi irodalomban egyre több spontánbeszéd-adatbázissal találkozni, mind a nemzetközi, mind a hazai spontánbeszéd-vizsgálatokra jellemző, hogy azok alapvetően fonetikai, illetve akusztikus sajátságok elemzésére irányulnak (pl. (Kane és mtsai, 2011; Reichel és Mády, 2013; Deme és Markó, 2013; Lennes és mtsai, 2009; Zhu és Penn, 2006)). Ugyanakkor azt, hogy olyan szemantikai–pragmatikai sajátságokról, mint például bizonyos beszédaktusok, illetve a nyelvi udvariasság különböző formái, pontos és valós képet kaphassunk, nagy méretű, megfelelően annotált spontánbeszéd-korpuszokra van feltétlenül szükség.

A nemzetközi korpuszok közül a legtöbb, amely pragmatikai annotációt is tartalmaz, telefonbeszélgetésekből áll. Így például a brit Telecom 1200 telefonbeszélgetéséből készült OASIS korpusz beszédaktus-szintű annotációt tartalmaz (Leech és mtsai, 2003). A Switchboard korpusz, amelyet több különböző sajátság mentén is annotáltak, szintén tartalmazza a beszédaktusok tagjeit is (Calhoun és mtsai, 2010). A dialógusaktusok jelentősen több típusát annotálták a fentebbi Switchboard korpusz egy részén (Jurafsky és mtsai, 1997).

Ami a magyar nyelvet illeti, jelenleg egyetlen olyan magyar, beszélt nyelvi korpuszról van tudomásunk (HuComTech), amely diskurzusszintű annotációt is tartalmaz, azonban ez az annotáció mindössze négy sajátságra terjed ki (turn-taking, turn-giving, backchannel, turn-keeping) (Pápay és mtsai, 2011).

2.4. Bizonytalanságra annotált korpuszok

A bizonytalanságot jelző nyelvi elemek vizsgálata intenzív kutatási területnek számít a számítógépes nyelvészetben, meg kell jegyeznünk ugyanakkor, hogy az eddigi vizsgálatok néhány kivételtől eltekintve az angol nyelv köré csoportosulnak, és elsődlegesen újsághíreket, biológiai publikációkat vagy orvosi dokumentumokat, illetve Wikipedia-szócikkeket elemeznek (vö. Szarvas és mtsai (2012); Kim és mtsai (2008); Saurí és Pustejovsky (2009)). Tudomásunk van mindemellett két magyar nyelvű, bizonytalanságra annotált korpuszról: a hUnCertainty korpusz magyar nyelvű Wikipédia-szócikkeket és bűnügyi híreket tartalmaz (Vincze, 2014), Vincze (2016) pedig közösségi médiából származó szövegekben foglalkozik bizonytalanság azonosításával. Az utóbb említett két munka annotációs sémáját alkalmaztuk mi is ebben a kutatásban.

3. A StaffTalk korpusz létrehozása

A StaffTalk korpusz hétköznapi szituációkban, spontán módon létrejött nyelvi tartalmakból áll, amelyek külső hatásoknak is kitett munkahelyi környezetben keletkeztek. A korpusz ezáltal lehetővé teszi a pletyka természetének valós, munkahelyi helyzetekben történő mélyebb megértését. A korpuszt spontán nyelvi produktumok alkotják, vagyis a kutatásban résztvevők szabadon megválaszthatták beszélgetésük tárgyát, hosszát és partnereit.

A beszélgetések rögzítése egy magyarországi iskola épületében zajlott 27 munkanapon keresztül. Az adatfelvétel 2019. április 8. és május 17. között zajlott (egy kisebb megszakítással, amikor iskolai program miatt pár nap kimaradt). Az adatfelvétel során a tanári közösség által legsűrűbben használt térre, a tanári szobára fókuszáltunk. A tanári közösség azon tagjai (összesen 20 fő), akik önként vállalták a kutatásban való részvételt, egy okosórát viseltek, mellyel rögzítettük a beszélgetéseiket. Az okosóra típusa Huawei SmartWatch 1 volt.[1] A rögzítő eszközt viselő személyek a nyakukban lévő figyelmeztető felirattal jelezték a rögzítés tényét a környezetükben lévő, kutatásban részt nem vevő személyek számára, akik szintén nyilatkoztak arról, hogy hozzájárulnak a hangfelvételeken történő szerepléshez. A hangrögzítés kizárólag abban az esetben indult el, amennyiben két eszköz megfelelő közelségbe került egymással és bármikor megállítható volt. Az etikai előírásoknak megfelelően, ha valaki úgy gondolta utólag, hogy az elhangzott beszélgetést mégsem szeretné rögzíteni, akkor azt utólag is jelezhette, természetesen ez esetben a már felvett hanganyagot töröltük. Az órák összesen 215:26:18 időtartamú hanganyagot rögzítettek.

A résztvevő tanárok többsége (70%) nő volt, átlagéletkoruk pedig 47 év (szórás: 9 év). 20 százalékuk rendelkezett a tanári munkáján felül valamilyen egyéb kiemelt beosztással: az intézményvezető, egy intézményvezető-helyettes, valamint két munkaközösség-vezető egyezett bele a részvételbe.

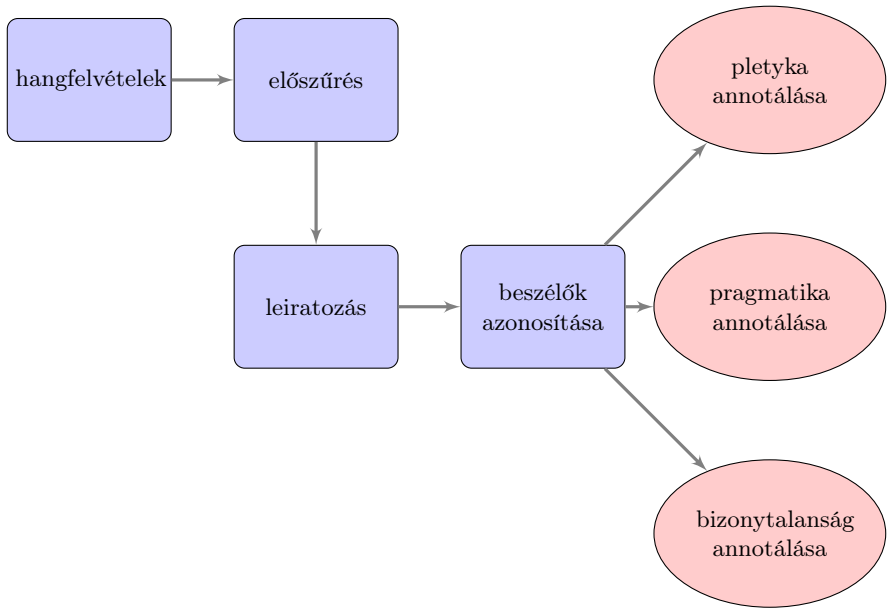
A projekt előkészítő szakaszában első lépésként a hangfájlokból kivágtuk a tíz másodpercnél hosszabb csendeket. Ennek eredményeként egy összesen 154:14:32 időtartamú hanganyag jött létre. Mielőtt a leiratozás megkezdődött, a hangfájlokat előválogattuk, amelynek során kiszűrtük a kutatás szempontjából nem releváns, adatvédelmi szempontból problémás, valamint nagyon rossz minőségű fájlokat. Amennyiben ugyanis például az okosórát viselő tanár elfelejtette az óráját kikapcsolni, az természetesen akár tanórákat, telefonbeszélgetéseket vagy diákokkal és szüleikkel történő beszélgetéseket is felvehetett, azonban ezeket kizártuk a feldolgozandó anyagok köréből. Nem foglalkoztunk az iskolai ünnepségek, tanári értekezletek és a kutatókkal való egyeztetések felvételeinek a feldolgozásával sem. A csak háttérzajt tartalmazó vagy nem megfelelő hangminőségű, leiratozásra alkalmatlan felvételeket ugyancsak kivettük a végleges korpuszból. Az előválogatás után 101:07:49 időtartamú hanganyag maradt (közel 47%-a az eredeti felvételeknek), a feldolgozás során ennek leiratozása, majd annotálása történt meg.

Ami az anyagok spontaneitását illeti, bár a résztvevők tudatában voltak annak, hogy beszélgetéseikről hangfelvételek készülnek, a folyamatos és hosszabb időn át tartó rögzítésnek köszönhetően spontán beszédnek tekinthetjük a felvett beszédanyagokat.

[1] Rögzítési frekvencia: 16kHz; csatornák: 1 (mono); bitmélység: 16 bit; nyers formátum: PCM; utófeldolgozás: az emberi hangos felerősítése, valamint a halk részek, végül a teljes anyag hangosítása.

4. A korpusz feldolgozása

Ebben a részben részletesen bemutatjuk a hanganyagok feldolgozásának teljes folyamatát, melyet az 1. ábra szemléltet.



1. ábra: A munkafolyamat.

4.1. A hanganyag leiratozása

Az előzetesen kiválogatott hangfelvételekhez – a további feldolgozást és annotációt megkönnyítendő – első lépésben leiratok készültek.

Az iskolai környezetből adódóan a felvételek némelyike zajos, nehezebben érthető. Sokszor többszereplős beszélgetéseket is tartalmaz a hanganyag, ahol a szereplők gyakran egymás szavába vágtak, egyszerre beszéltek. A jelenleg a magyar nyelvhez rendelkezésre álló automatikus beszédfelismerő alkalmazások e jelenségek többségére nincsenek felkészítve, így ezek használatát elvetettük, és a kézi leiratozás mellett döntöttünk.

Ebben a fázisban tíz gépelő vett részt, akik a hallott anyagot legépelték, időbélyegekkel, illetve különféle annotációkkal látták el. Elkülönítették egymástól az egyes beszélők megszólalásait, valamint bizonyos hanghatásokat és beszédjellemzőket (pl. suttogás, nevetés) is jelöltek. Ezeken felül az egyszerre beszélést, egymás szavába vágást is jelölték, valamint a nem és nem biztosan jól értett részeket is. Végezetül az egyes beszélgetések határait is annotálták a leiratok-

ban. A munkafolyamat során az F4 szoftvert használták[2]. Egy órányi hanganyag leiratozása átlagosan 14 órájába telt a gépelőknek.

A leiratozás minőségének biztosítása érdekében rendszeres időközönként bizonyos hangfájlokat (összesen a korpusz közel 10%-át) több gépelővel is leírattunk, majd az így kapott szövegváltozatokat automatikusan összevetettük egymással. Vizsgáltuk a szókincs egyezését, valamint a szövegben elhelyezett annotáció minőségét és az egyes tagek mennyiségét. Ha az egyezés mértéke nem felelt meg az előzetes elvárásoknak (60%), akkor a fájlokat további javításra visszaadtuk az adott leiratozónak, ez körülbelül az ellenőrzött fájlok 15%-át érintette.

4.2. A beszélők azonosítása

Fő kutatási kérdéseink egyike volt, hogy egy zárt csoporton belül hogyan befolyásolja a kommunikáció, különösen a pletyka a közösség működését. Ennek vizsgálatához elengedhetetlen, hogy rendelkezésünkre álljon az adott beszélgetésben részt vevő személyek kiléte is, azaz tudjuk, ki, mikor, kivel és miről beszélgetett. Elengedhetetlen volt tehát az egyes diskurzusokban részt vevő személyek név szerinti azonosítása is.

A kutatás 20 résztvevőjéről rendelkezésre álltak hangminták is, illetve ismert volt az az információ is, hogy melyik órát ki viseli. Ugyanakkor a résztvevők beszélgethettek olyan személyekkel is, akik nem vettek részt az adatfelvételben (pl. diákok, szülők), így a hangfelvételeken megszólaló személyek száma meghaladja a 30-40-et is. Mivel előzetes tapasztalataink azt mutatták, hogy a gépelést jelentősen lelassítja, ha a leiratozónak kell egyúttal azonosítani is a beszélőket, külön munkafázisba szerveztük ezt a feladatot: két erre szakosodott nyelvész (kutatói támogatással) végezte a már elkészített leiratok alapján az egyes megszólalók azonosítását. A kutatásban részt nem vevők hangját külön azonosítóval („külső személy") láttuk el.

A fentebbi munkaszervezési megoldással nagymértékben sikerült meggyorsítani a leiratozás folyamatát, valamint hatékonyabbá tenni a beszélők azonosítását.

E fázist követően a létrejött szövegfájlokat három különálló fázisban annotáltuk, amelyhez az MMAX2 eszközt (Müller és Strube, 2006) használtuk. Az annotátorok időnként ugyanazokat a fájlokat annotálták egymástól függetlenül, így munkájuk minőségét össze tudtuk vetni. A minőségbiztosítás fájljainak mennyiségét úgy határoztuk meg, hogy kitegye a teljes korpusz 10%-át, és ezáltal megfeleljen a nemzetközi sztenderdnek. Annak céljából pedig, hogy az ellenőrzés folyamatos lehessen, ezeket a fájlokat a munka teljes hosszában arányosan osztottuk ki. Amennyiben az annotátorok közti egyetértés nem érte el a kívánt szintet, a fájlokat utólag ki kellett javítaniuk, erre azonban az eseteknek csupán a töredékében (körülbelül 15%-ában) volt szükség, és az annotátoroknak minden esetben sikerült a problémás fájlokat megfelelően korrigálni.

Az egyes fázisok részleteit az alábbiakban ismertetjük.

[2] https://www.audiotranskription.de/english/f4

4.3. Pletyka

A *pletyka* annak hagyományos definíciója szerint olyan, jelen nem lévő személyről vagy személyekről folytatott értékelő tartalmat hordozó beszélgetéseket takar, amelyekben azt értékelést megfogalmazó személy és legalább egy hallgató jelen van (Emler, 1994; Grosser és mtsai, 2012). A *pletyka* fogalma gyakran negatív konnotációt hordoz, azonban fontos szerepet tölthet be az információáramlásban, a személyek közti kapcsolatok megerősítésében, a csoportnormák fenntartásában és betartatásában vagy szelepként szolgálhat a felgyülemlett negatív érzelmek „kiengedésében" (Grosser és mtsai, 2012).

A jelen tanulmányban bemutatott korpusz építése során egyéb, személyekre vagy a köztük lévő kapcsolatra irányuló beszédtartalmakat is vizsgálunk. A személyekre vonatkozó beszédtartalmak annotálását két fő dimenzió mentén végeztük, a pletyka célszemélye vagy célszemélyei tekintetében. Csoporton belüli pletykának tekintettük az annotálás során, amennyiben a pletyka célszemélye a munkatársi csoporthoz tartozó személy, míg csoporton kívüli pletykának, amennyiben a célszemély nem a munkatársi csoporthoz tartozik. A munkatársi csoporthoz tartozó és nem tartozó személyekről szóló beszédtartalom megkülönböztetése azért releváns, mert eltérő funkciót tölthetnek a szervezeten belüli kommunikációban.

Jelenlévő célszemélyre vonatkozó beszédtartalomnak tekintettük, amennyiben az összes, a pletykában említett célszemély jelen volt a beszélgetésben, míg jelen nem lévő célszemélyre vonatkozó beszédtartalomnak, amennyiben az említett személy vagy személyek nem voltak jelen a beszélgetésben. A munkatársi csoporthoz tartozó és nem tartozó, valamint a jelenlévő és jelen nem lévő célszemélyekre vonatkozó pletyka vegyesen is előfordulhat egy pletykán belül (főként személyek közti relációra vonatkozó tartalmak esetén), így ezeket is megkülönböztettük. A két dimenzió mentén való besoroláson felül pedig további ismérvek jelölését is kértük az annotátoroktól a személyekre vonatkozó beszédtartalmakra nézve, az alábbiaknak megfelelően:

- típus: csoporton belüli vagy kívüli személyekre, illetve jelen lévő vagy jelen nem lévő személyekre irányul a pletyka
- polaritás: pozitív, negatív vagy semleges a szövegtartalom
- forrás: a pletyka közlője
- célpont: az a személy, akiről szól a pletyka
- reláció: két vagy több személy relációjáról szól-e a pletyka
- normativitás: normatív magatartással függ-e össze a pletyka

Jelölték ezen felül, amennyiben egy pletykára a pletykát megerősítő, továbbvivő, illetve azt hárító reakció érkezett. Amennyiben a pletykára érkező reakció egyben új pletykát is tartalmazott, akkor az adott szövegrészt pletykaként és reakcióként is annotálták.

Több pletyka összefűzhető volt egy láncba, amennyiben a pletyka ugyanazon témában ugyanazon célszemélyre vonatkozott.

A fentebb bemutatottakon túl a korpusz egyéb információk kinyerését is lehetővé teszi, amelyek annotálására nem volt szükség. Így például, mivel a hangok

azonosítása is megtörtént, akár a munkahelyi hierarchia, illetve az életkor és a pletyka összefüggései is vizsgálhatóak lesznek a korpusz segítségével a jövőben.

4.4. Pragmatika

A közösségen belüli kommunikáció vizsgálatának egyik fontos vetülete, hogy milyen beszédaktusokat és udvariassági stratégiákat használnak egymás között az egyes közösségi tagok. Ennek vizsgálatához részletes pragmatikai annotációval láttuk el a leiratokat, az alábbi annotációs sémát alkalmazva.

- Beszédaktusok:
  - ígéret / ajánlat
  - figyelmeztetés / fenyegetés
  - kérés / parancs / kívánság
  - panasz / vád / kritika / sértés
  - dicséret / bók
  - bocsánatkérés
  - köszönetnyilvánítás
- Reakciók:
  - elfogadás / egyetértés
  - visszautasítás / egyet nem értés
  - hárítás (nem egyértelmű elfogadás vagy visszautasítás)
- Irónia:
  - irónia
  - antiirónia
- Interakciós elemek:
  - figyelem felhívása
  - üdvözlés / elköszönés

A pragmatikaihoz annotációt két nyelvész készítette a már említett MMAX2 szoftver (Müller és Strube, 2006) segítségével. A pragmatikai (és bizonytalansági) annotációkat részletesen taglaljuk egy másik, az MSZNY2021 konferencián megjelent cikkben (Vincze és mtsai, 2021).

4.5. Bizonytalanság

Harmadik annotációs szintként az annotátorok megjelölték a nyelvi bizonytalanságra utaló szavakat a korpuszban. Úgy gondoljuk, hogy a bizonytalanság annotálása összeköthető mindkét másik annotációs szinttel: egyfelől feltételezzük, hogy a pletyka közlője bizonytalanságot hordozó nyelvi elemeket is beleszőhet a mondandójába, ami a pletykában az egyes típusok gyakoriságát illetően egyfajta, eleddig ismeretlen mintázatot mutathat. Másfelől bizonyos beszédaktusok és a bizonytalanság kifejezőeszközei sokszor egybeesnek (például kérésekben gyakran szerepel feltételes módú ige), így érdemes összevetni ezen nyelvi elemek többféle szerepét ugyanabban a nyelvi adatbázisban.

A bizonytalanság annotálásakor követtük a már korábban létrehozott magyar nyelvű bizonytalansági korpuszok kategorizálását (Vincze, 2014, 2016), amelyet az alábbiakban foglalunk össze:

- Szemantikus bizonytalanság:
  - episztemikus
  - doxasztikus
  - feltételes
  - vizsgálat
- Diskurzusszintű bizonytalanság:
  - weasel: bizonytalan információforrás vagy szereplő a cselekvésben
  - hedge: mennyiségek vagy minőségek homályos jelölése
  - peacock: bizonyít(hat)atlan állítás vagy túlzás

A bizonytalanság annotálását – a pragmatikaihoz hasonlóan – két nyelvész végezte az MMAX2 szoftver (Müller és Strube, 2006) segítségével. Ahogy már fentebb említettük, a bizonytalansági annotációkat is részletesen elemezzük egy másik, az MSZNY2021 konferencián megjelent cikkben (Vincze és mtsai, 2021).

5. Statisztikai adatok

A korpuszban található annotációk mennyiségi megoszlása az 1. táblázatban látható.

Annotációs szint	Hanganyag időtartama	Mondatszám	Tokenszám	Annotált egységek száma
Pletyka	102:42:30	124 836	1 461 769	44 165
Pragmatika	102:42:30	124 836	1 461 769	26 463
Bizonytalanság	102:42:30	124 836	1 461 769	28 340

1. táblázat. A korpusz adatai.

6. A korpusz felhasználhatósága

A StaffTalk korpusz – kézi leiratozása és részletes annotációja, továbbá keletkezési körülményei miatt – egyaránt hasznos lehet mind a beszédtechnológusoknak, mind számítógépes nyelvészeknek, elméleti és alkalmazott nyelvészeknek, valamint a társadalomtudósoknak.

A korpuszépítés befejezését követően a korpuszt adatvédelmi okok miatt anonimizáljuk. Az anonimizálást követően a korpuszt bárki számára kutatási és oktatási célra elérhetővé tesszük.

7. Összegzés

A dolgozatban bemutattunk a StaffTalk korpuszt, amely a magyar nyelvű spontán diskurzust reprezentálja, a hangzó szövegek legépelt és annotált változatával együtt. A legépelt hanganyagokat három szinten annotáltuk: a beszélgetésekben

megjelenő pletykát, a beszédaktusokat és egyéb pragmatikai jegyeket, valamint a bizonytalanságra utaló szavakat egyaránt megjelöltük. A tanulmány célja az volt, hogy részletes információval szolgáljon a korpusz készítésének céljáról, eszközeiről és módjáról, valamint ismertesse annak alapvető statisztikai adatait.

A korpuszban foglalt szöveganyag, valamint a feldolgozás módja teret nyit számos olyan vizsgálat elvégzésére a jövőben, amely a spontán beszéd természetét kívánja kutatni, válaszokat adva ezzel bizonyos, a humán kommunikáció, illetve interakció természetét érintő kérdésekre. Mind a pletykát érintő, mind az udvariasság és bizonytalanság témakörébe tartozó kutatási kérdéseinket szeretnénk behatóan vizsgálni és tárgyalni a jövőben a korpusz segítségével.

A hangrögzítést kiegészítette egy napi szintű kérdőíves felmérés, mely a kutatásban résztvevők informális és formális kapcsolataira, munkahelyi elégedettségére és közérzetére vonatkozó kérdéseket tartalmazott. A kiegészítő adatok lehetővé teszik a pletyka más hálózatokkal való összefüggésben történő vizsgálatát is.

Tervezzük a korpusz nyilvánossá tételét a jövőben a kutatók számára, az adatok anonimizálását követően.

Köszönetnyilvánítás

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

Szabó Martina Katalin kutatásait részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal – NKFIH OTKA posztdoktori kiválósági programja (NKFI-azonosító: 132312) támogatta.

Szeretnénk köszönetet mondani a korpusz leiratozásában és annotálásában tevékenyen részt vállaló kollégáinknak kitartó és lelkes munkájukért.

Hivatkozások

Babarczy, A.: Analógikus általánosítási folyamatok a gyereknyelvben= analogical generalisation processes in language acquisition. OTKA Kutatási Jelentések| OTKA Research Reports (2009)

Bodó, Cs., Kocsis, Zs., Vargha, F.: A Budapesti Egyetemi Kollégiumi Korpusz. Elméleti és módszertani kérdések. In: Benő, A., Fazakas, N. (szerk.) Élőnyelvi kutatások és a dialektológia: Válogatás a 19. Élőnyelvi Konferencia - Marosvásárhely, 2016. szeptember 7-9. - előadásaiból. pp. 169–177 (2017)

Calhoun, S., Carletta, J., Brenier, J.M., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D.: The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language resources and evaluation 44(4), 387–419 (2010)

Crowdy, S.: Spoken corpus design. Literary and Linguistic Computing 8(4), 259–265 (1993)

Deme, A., Markó, A.: Lengthenings and filled pauses in Hungarian adults' and children's speech. In: Proceedings of DiSS 2013, The 6th Workshop on Disfluency in Spontaneous Speech. TMH-QPSR 54:1. vol. 54, pp. 21–24. KTH Royal Institute of Technology (2013)

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

Dunbar, R.I.: Gossip in evolutionary perspective. Review of General Psychology 8(2), 100–110 (2004)

Eckhaus, E., Ben-Hador, B.: Gossip and gender differences: a content analysis approach. Journal of Gender Studies 28(1), 97–108 (2019)

Emler, N.: Gossip, reputation and social adaptation. In: Goodman, R.F., Ben-Ze'ev, A. (szerk.) Good gossip. University Press of Kansas, Lawrence (1994)

Foster, E.K.: Research on gossip: Taxonomy, methods, and future directions. Review of general psychology 8(2), 78–99 (2004)

Galántai, J., Pápay, B., Kubik, B.G., Szabó, M.K., Takács, K.: A pletyka a társas rend szolgálatában-az informális kommunikáció struktúrájának mélyebb megértéséért a computational social science eszközeivel. Magyar Tudomány 179(7), 964–976 (2018)

Gósy, M.: BEA–A multifunctional Hungarian spoken language database. Phonetician 105, 50–61 (2013)

Grosser, T., Kidwell, V., Labianca, G.J.: Hearing it through the grapevine: Positive and negative workplace gossip. Organizational Dynamics 41, 52–61 (2012)

Gulyás, A., Galántai, J., Szabó, M.K., Szebeni, Z.: A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 317–330. Szegedi Tudományegyetem, Szeged (2018)

Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The ATIS spoken language systems pilot corpus. In: Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990 (1990)

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Van Ess-Dykema, C.: Automatic detection of discourse structure for speech recognition and understanding. In: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings. pp. 88–95. IEEE (1997)

Kane, J., Pápay, K., Hunyadi, L., Gobl, C.: On the Use of Creak in Hungarian Spontaneous Speech. In: ICPhS. pp. 1014–1017 (2011)

Kim, J.D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 9(Suppl 10) (2008), http://www.biomedcentral.com/1471-2105/9/10

Kugler, N.: Megfigyelés és következtetés a nyelvi tevékenységben. No. 179, Tinta Könyvkiadó (2015)

Leech, G., McEnery, T., Weisser, M.: Spaac speech-act annotation scheme. University of Lancaster (2003)

Lennes, M., és mtsai: Segmental features in spontaneous and read-aloud finnish. Phonetics of Russian and Finnish general description of phonetic systems: experimental studies on spontaneous and read-aloud speech (2009)

Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous Speech Corpus of Japanese. In: LREC. pp. 947–9520. Citeseer (2000)
McEnery, T.: Corpus linguistics, vol. 978019. Oxford University Press Inc (2012)
Mengusoglu, E., Deroo, O.: Turkish lvcsr: Database preparation and language modeling for an agglutinative language. In: IEEE International Conference on Acoustics Speech And Signal Processing. vol. 6, pp. 4018–4018. IEEE; 1999 (2001)
Mitra, T., Gilbert, E.: Have you heard?: How gossip flows through workplace email. In: ICWSM (2012)
Müller, C., Strube, M.: Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (szerk.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 197–214. Peter Lang, Frankfurt a.M., Germany (2006)
Neuberger, T., Gyarmathy, D., Gráczi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: International Conference on Text, Speech, and Dialogue. pp. 424–431. Springer (2014)
Oostdijk, N.: The spoken dutch corpus. overview and first evaluation. In: LREC. pp. 887–894. Athens, Greece (2000)
Pápay, K., Szeghalmy, S., Szekrényes, I.: Hucomtech multimodal corpus annotation. Argumentum 7, 330–347 (2011)
Reichel, U.D., Mády, K.: Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech. In: Wagner, P. (szerk.) Elektronische Sprachsignalverarbeitung 2013. pp. 223–230. TUDpress, Dresden (2013), http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-18043-4
Robbins, M.L., Karan, A.: Who gossips and how in everyday life? Social Psychological and Personality Science 11(2), 185–195 (2020)
Saurí, R., Pustejovsky, J.: FactBank: a corpus annotated with event factuality. Language Resources and Evaluation 43, 227–268 (2009), http://dx.doi.org/10.1007/s10579-009-9089-9
Seppänen, T., Toivanen, J., Väyrynen, E.: MediaTeam speech corpus: a first large Finnish emotional speech database. In: Proceedings of the Proceedings of XV International Conference of Phonetic Science. pp. 2469–2472. Citeseer (2003)
Szabó, M.K., Galántai, J.: Egy magyar nyelvű spontán beszélt nyelvi korpusz (HuTongue) létrehozásának tapasztalatai. In: XXVI. MANYE Kongresszus konferenciakötete. Pécs (2017)
Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., Gurevych, I.: Cross-genre and cross-domain detection of semantic uncertainty. Computational Linguistics 38, 335–367 (June 2012)
Van Bael, C., Baayen, R.H., Strik, H.: Segment deletion in spontaneous speech: a corpus study using mixed effects models with crossed random effects. In: INTERSPEECH. pp. 2741–2744 (2007)
Váradi, T.: A budapesti szociolingvisztikai interjú. In: Kiefer, F., Siptár, P. (szerk.) A magyar nyelv kézikönyve, pp. 339–359. Akadémiai Könyvkiadó, Budapest (2003)

Vicsi, K., Tóth, L., Kocsor, A., Csirik, J.: MTBA–a Hungarian telephone speech database. Híradástechnika, LVII 8 (2002)

Vincze, V.: Uncertainty detection in Hungarian texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1844–1853. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014), https://www.aclweb.org/anthology/C14-1174

Vincze, V.: Detecting uncertainty cues in Hungarian social media texts. In: Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM). pp. 11–21. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://www.aclweb.org/anthology/W16-5002

Vincze, V., Üveges, I., Szabó, M.K.: Magyar nyelvű spontán beszéd szemantikai–pragmatikai sajátságainak elemzése nagy méretű korpusz (StaffTalk) alapján. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)

Zhu, X., Penn, G.: Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. pp. 197–200 (2006)