Magyar nyelvű spontán beszéd szemantikai–pragmatikai sajátságainak elemzése nagy méretű korpusz (StaffTalk) alapján

Vincze Veronika[1], Üveges István[2,3], Szabó Martina Katalin[2,4]

[1]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.
[2]Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.
[3]Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
6722 Szeged, Egyetem utca 2.
[4]Társadalomtudományi Kutatóközpont, CSS-RECENS
1097 Budapest, Tóth Kálmán utca 4.
{vinczev,martina}@inf.u-szeged.hu, uvegesistvan898@gmail.com

Kivonat A dolgozatban bizonyos pragmatikai és szemantikai sajátságokat vizsgálunk magyar nyelvű, nagy méretű spontánbeszéd-korpusz (StaffTalk) alapján. A vizsgálati korpusz is hétköznapi szituációkban, külső hatásoknak is kitett munkahelyi környezetben, spontán módon létrejött nyelvi tartalmakból áll, vagyis a kutatásban résztvevők szabadon megválaszthatták beszélgetésük tárgyát, hosszát és partnereit. A korpusz létrehozása során a hanganyagokat legépelték, majd azt követően számos szempont alapján annotálták. A jelen vizsgálatokat ezeket az annotációkat felhasználva végezzük el.
Kulcsszavak: korpusz, spontán beszéd, magyar, pragmatika, beszédaktusok, udvariasság, bizonytalanság

1. Bevezetés

A dolgozatban a magyar spontán beszéd bizonyos szemantikai és pragmatikai sajátságait vizsgáljuk nagy méretű, kézzel annotált korpusz alapján. A cikk hiánypótló, hiszen a magyar spontán beszéd bizonyos szemantikai és pragmatikai sajátságait vizsgálja egy egyedülálló, nagy méretű, kézzel annotált spontánbeszéd-korpusz alapján. A hiány oka alapvetően az, hogy még nemzetközi szinten is szerény azoknak az adatbázisoknak a száma, amelyek a spontán beszédet reprezentálják, illetve azoké is, amelyek valamilyen kézzel készített, szemantikai–pragmatikai annotációval rendelkeznek.

A vizsgálati korpusz, amely az első olyan, magyar nyelvű spontánbeszéd-adatbázis, amely számos szemantikai és pragmatikai sajátság kézi annotációját tartalmazza, hétköznapi szituációkban, spontán módon létrejött nyelvi tartalmakból áll, amelyek külső hatásoknak is kitett munkahelyi környezetben

keletkeztek. A beszélgetések rögzítése egy magyarországi iskola épületében zajlott 27 munkanapon keresztül. A tanári munkaközösség azon tagjai (összesen 20 fő), akik önként vállalták a kutatásban való részvételt, egy okosórát viseltek, mellyel felvették egymás közti beszélgetéseiket. A hangfájlokat egy annotátorcsapattal legépeltettük és három különböző fázisban annotáltattuk, amelyek a következők voltak: pletykadiskurzusok, bizonyos pragmatikai sajátságok, valamint bizonytalanságra utaló nyelvi elemek.

A jelen dolgozatban tett megállapításokhoz a két utóbbi szinten létrehozott annotációt használjuk fel. A közösségen belüli kommunikáció vizsgálatának egyik fontos vetülete, hogy milyen beszédaktusokat és udvariassági stratégiákat használnak egymás között az egyes közösségi tagok. Ugyanakkor a nyelvi bizonytalanság külön is figyelmet érdemel és kifejezőeszközei sokszor egybeesnek bizonyos beszédaktusokkal. Mindezek beható vizsgálatára, valamint összefüggéseinek feltárására ad lehetőséget a StaffTalk korpusz részletes annotációja. Az elvégzett korpuszvizsgálatok új eredményeket hozhatnak a spontán beszélt nyelv diskurzusaktusainak kvantitatív és kvalitatív sajátságairól, amelyek összevethetően a korábbi, elméleti szintű nyelvészeti megállapításokkal.

2. Kapcsolódó irodalom

Bár a nemzetközi irodalomban egyre több spontánbeszéd-adatbázissal találkozni (Crowdy, 1993; Hemphill és mtsai, 1990; Maekawa és mtsai, 2000; Oostdijk, 2000; Van Bael és mtsai, 2007; Neuberger és mtsai, 2014), a magyar nyelvű korpuszok száma ezen a téren messze elmarad a nemzetközitől. Ugyanakkor mind a nemzetközi, mind a hazai spontánbeszéd-vizsgálatokra egyaránt 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)).

A nemzetközi korpuszok közül a legtöbb, amely pragmatikai annotációt is tartalmaz, írott szövegekből (De Felice és mtsai, 2013) vagy, amennyiben hangzó szövegeket tartalmaz – telefonbeszélgetésekből készült (Leech és mtsai, 2003). 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. A munka célja az volt, hogy vizsgálati és tanító anyagot készítsenek a természetesnyelvi interakció statisztikai modellezésére és a diskurzusstruktúrák automatikus detektálására (Jurafsky és mtsai, 1997). Más, nem telefonbeszélgetéseket tartalmazó spontánbeszéd-korpuszok nemzetközi szinten is kifejezetten ritkák (pl. (Cheng és mtsai, 2005)).

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).

A kutatócsoportunk által készített másik nagyméretű korpusz, a HuTongue (Galántai és mtsai, 2018; Gulyás és mtsai, 2018) csupán félig (vagy részlegesen) spontánnak tekinthető, 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. Ugyanakkor olyan szemantikai–pragmatikai sajátságok, mint például bizonyos beszédaktusok, a nyelvi udvariasság különböző formái vagy a bizonytalanság beható vizsgálata csupán nagy méretű, kézzel megfelelően annotált spontánbeszéd-korpusz alapján lehetséges.

A fejezet további részében néhány sajátság kiemelésével szeretnénk jobban rámutatni az elkészített korpusz annotációjának a fontosságára.

Danescu-Niculescu-Mizil és mtsai (2013) alapján az udvariasság például a humán kommunikáció központi motorja, éppoly alapvető, mint az igazmondás, az informativitás, a relevancia vagy a világosság követelményei (Paul és mtsai, 1975; Leech, 2016; Brown és Levinson, 1978). A természetes nyelvben az udvariasság számtalan eszközzel és variációban kódolható (Danescu-Niculescu-Mizil és mtsai, 2013). Markerei szorosan kapcsolódnak a társadalmi interakciók hatalmi dinamikájához, és gyakran meghatározó tényezők abban, hogy ezek az interakciók jól vagy rosszul működnek-e (Andersson és Pearson, 1999; Rogers és Lee-Wong, 2003; Holmes és Stubbe, 2015).

A beszédaktusok kutatásának egyik központi kérdése a különbségtétel a lokúciós és az illokúciós jelentés között (Austin, 1975; De Felice és mtsai, 2013). Röviden összefoglalva, az előbbi a beszédaktus szó szerinti, kimondott vagy leírt tartalmára utal (pl. túl alacsony vagyok ahhoz, hogy elérjem a polcot), míg utóbbi a beszédaktus funkcióra utal, ami a beszélő a megnyilatkozásával valójában kommunikálni szándékozik (pl. segítség kérése a magas polcon lévő tárgy eléréséhez). Az, hogy hogyan lehet helytállóan elszámolni ezzel a különbséggel, a beszédaktus annotációjának egyik fő kihívása (De Felice és mtsai, 2013).

Az ún. indirekt beszédaktusok esetében, Searle (1975) definíciója alapján egy adott beszédaktust a beszélő egy másik beszédaktussal valósítja meg. E beszédaktusok megfelelő kezelésének kiemelt jelentősége van többek között a mesterséges intelligencia területén, hiszen ennek hiánya csökkenti az intelligens rendszerek hatékonyságát az emberi környezettel való interakcióban (Roque és mtsai, 2020).

3. Az annotált korpusz

A StaffTalk korpusz hétköznapi szituációkban, spontán módon létrejött nyelvi tartalmakból áll, amelyek külső hatásoknak is kitett munkahelyi környezetben keletkeztek 27 munkanap alatt. A korpuszt spontán nyelvi produktumok alkotják, vagyis a kutatásban résztvevők szabadon megválaszthatták beszélgetésük tárgyát, hosszát és partnereit. A résztvevők okosórát viseltek, melyek rögzítették beszélgetéseiket. (Mindezekről részletesebben beszámolunk egy másik, ugyanezen a konferencián megjelent dolgozatban (Szabó és mtsai, 2021)).

Az órák összesen 215:26:18 időtartamú hanganyagot rögzítettek. A projekt előkészítő szakaszában, első lépésként a hangfájlokból kivágtuk a tíz másodpercnél hosszabb csendeket, majd az anyagot tovább válogattuk: kiszűrtük a kutatás szempontjából nem releváns, adatvédelmi szempontból problémás, valamint nagyon rossz minőségű fájlokat. Az előválogatás után 105:16:10 időtartamú hanganyag maradt (közel 47%-a az eredeti felvételeknek), a feldolgozás során ennek leiratozása, majd annotálása történt meg.

A leiratozási fázisban tíz gépelő vett részt, akik a hallott anyagot legépelték, időbélyegekkel, illetve különféle annotációkkal látták el. Elengedhetetlen volt az egyes diskurzusokban részt vevő személyek név szerinti azonosítása is. (Mindezekről ugyancsak részletesen beszámol Szabó és mtsai (2021).) E fázist követően a létrejött szövegfájlokat három különálló fázisban annotáltattuk, amelyhez az MMAX2 eszközt (Müller és Strube, 2006) használtuk. A munka során a pletykadiskurzusokat, különböző pragmatikai sajátságokat, valamint a nyelvi bizonytalanság jelölőit annotáltattuk sokrétűen.

Ebben a fejezetben részletesen bemutatjuk a korpusz két, pletykán kívüli annotációját, valamint közlünk néhány megállapítást az annotáció alapján végzett statisztikai vizsgálatokról.

3.1. Pragmatika

A különböző pragmatikai jelenségek esetében – amennyiben lehetséges volt – (minimum) teljes tagmondatokat jelöltünk. Ha egymás után több tagmondat/mondat is ugyanabba a kategóriába tartozott (pl. hosszasan panaszkodott valaki), akkor azt egy egységként jelöltük be.

A pragmatikai egységek típusát illetően egyaránt figyelembe vettük az Austin-Searle neve által fémjelzett beszédaktuselméletet (Austin, 1975; Searle, 1975), udvariasságelméleteket (Brown és Levinson, 1978), valamint az ezekre adott lehetséges reakciókat, valamint az irónia és antiirónia jelenségeit. Különálló kategóriaként vettük fel a „figyelem felhívása" beszédaktust, mivel úgy véljük, hogy spontán beszélgetésekben ennek kiemelt szerepe és gyakorisága lehet, a beszédpartnerek személyes interakciójának köszönhetően. Hangsúlyoznunk kell azt is, hogy több, hagyományosan különállónak tekintett beszédaktust összevontunk a jelen annotációs sémában, elsősorban azért, mert pusztán a leírt és hallott beszédre támaszkodva nem kaphatunk teljes képet a beszélő motivációiról, szándékairól, érzelmeinek erősségéről, ami például a figyelmeztetés és fenyegetés elkülönítésében kulcsfontosságú szerepet kapna.

A nyelvi bizonytalanság annotálásakor azt a minimális egységet/kulcsszót (szót vagy szókapcsolatot) jelöltük, amely önmagában is felelős volt a bizonytalanságért, pl. *talán, lehet, szerintem, nem is tudom* stb. Ható és feltételes módú igék esetében, amennyiben bizonytalan jelentéstartalommal rendelkeztek, a teljes igét jelöltük (mivel morfémát nem tudtuk önmagában kijelölni). Több szó együtt tehát kizárólag akkor volt jelölhető, ha együtt hordozta a bizonytalan tartalmat (pl. *tudom* vs. *nem tudom*).

Ami a nyelvi bizonytalanság típusait illeti, sok esetben valamely lexikális tartalom, másképpen egy konkrét nyelvi elem felelős a bizonytalanságért egyfaj-

ta bizonytalansági markerként. Más típusú bizonytalanságok esetében azonban nem lehet csupán a szemantikára koncentrálni, ugyanis a bizonytalanságot a ko-, illetve kontextus határozza meg. Az előbbit a fentebbieknek megfelelően szemantikai, az utóbbit diskurzusszintű bizonytalanságnak nevezzük, és azoknak több altípusát különböztetjük meg (Vincze, 2013).

Mindkét annotációs szint esetében azt kértük az annotátoroktól, hogy a munkát a hanganyag hallgatásával egyszerre végezzék, és a jelöléseket mindig az aktuális kontextus és hangsúly, illetve hanglejtés függvényében végezzék el.

A beszédaktusokat és pragmatikai sajátságokat az alábbi annotációs séma szerint annotáltuk. (A kevésbé egyértelműekhez rövid magyarázatot fűzünk.)

- Beszédaktusok:
  - ígéret / ajánlat (jövőbeli pozitív cselekedetre utalás)
  - figyelmeztetés / fenyegetés (jövőbeli negatív cselekedetre utalás)
  - kérés / parancs / kívánság
  - panasz / vád / kritika / sértés (a partner vagy bármely személy (a beszélő maga is lehet) iránti negatív vélemény kifejezése negatív jelentéstartalmú szavakkal)
  - dicséret / bók (a partner vagy bármely személy (a beszélő maga is lehet) iránti pozitív vélemény kifejezése pozitív jelentéstartalmú szavakkal)
  - bocsánatkérés
  - köszönetnyilvánítás
- Reakciók:
  - elfogadás / egyetértés
  - visszautasítás / egyet nem értés (nyílt visszautasítás/egyet nem értés; ajánlatra, kérésre adott direkt visszautasítás vagy az egyet nem értés nyílt kifejezése)
  - hárítás (ajánlatra, kérésre adott válaszként, nem derül ki, hogy egyetért vagy nem ért egyet az előzőekkel, tehát ez az egyet nem értés vagy visszautasítás „kikerülése")
- Irónia:
  - irónia (a beszélői szándék szerint a szótári jelentéssel ellentétes értékkel használt szavakkal kifejezett megnyilatkozások, tehát pozitív jelentéstartalmú szavakkal kifejezett negatív tartalom)
  - antiirónia (negatív jelentéstartalmú szavakkal kifejezett pozitív értékelés)
- Interakciós elemek:
  - figyelem felhívása (fontos vagy érdekes mondandó jelzése a partner felé)
  - üdvözlés / elköszönés

A pragmatikai annotációt képzett nyelvészek végezték az MMAX2 szoftver (Müller és Strube, 2006) segítségével.

3.2. Bizonytalanság

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

– Szemantikus bizonytalanság:
  - episztemikus: a világtudásunk alapján nem tudjuk eldönteni, hogy igaz-e vagy hamis az állítás. Gyakran ható igékkel fejeződik ki, de más lexikai elemek is előfordulhatnak (*talán, valószínűleg, lehetséges*)
  - doxasztikus: hiedelmek, vélemény kifejezése (*hisz, gondol, vél, szerint*)
  - feltételes: egy adott feltételhez kötött az állítás igazságértéke (*ha... akkor*)
  - vizsgálat: pl. kutatási kérdés egy tudományos cikkben (*megvizsgál, elemez*)
– Diskurzusszintű bizonytalanság:
  - weasel: bizonytalan információforrás vagy szereplő a cselekvésben (*valaki, egyesek*)
  - hedge: mennyiségek vagy minőségek homályos jelölése (*sok, gyakori*)
  - peacock: bizonyít(hat)atlan állítás vagy túlzás (*gyönyörűszép, botrányos*)

A bizonytalanság annotálását – a pragmatikaihoz hasonlóan – képzett nyelvészek végezték az MMAX2 szoftver (Müller és Strube, 2006) segítségével.

4. Eredmények

Ebben a fejezetben összegezzük a kétféle annotációs szint eredményeit, valamint röviden elemezzük a pragmatikai és bizonytalansági annotáció kapcsolatát.

4.1. Pragmatikai annotáció

Az annotált korpuszban található pragmatikai annotált elemek gyakoriságát az 1. táblázat szemlélteti.

Az adatokból kiviláglik, hogy a leggyakoribb kategória az elfogadás / egyetértés, azaz a beszélgetőpartnerek leginkább helyeslésüket fejezték ki a másik mondandója iránt. Ennek interakciós párja, a visszautasítás / egyet nem értés ugyanakkor ennél ritkábban fordul elő a korpuszban, a hatodik helyen található. Érdekes ugyanakkor megfigyelni, hogy a hárítás csak a tizenkettedik helyen szerepel, azaz a beszélgetőpartnerek inkább nyíltan felvállalják egyet nem értésüket, mintsem hogy kikerüljék a véleménynyilvánítást. Ezt valószínűleg magyarázhatja azt is, hogy zárt közösségben készültek a hangfelvételek, a partnerek jól ismerik egymást, kicsi a köztük levő szociális távolság, ami együtt jár az udvariassági stratégiák alkalmazásának csökkenésével (Wolfson, 1988).

A 3-4., valamint a 7-8. leggyakoribb kategóriák is pozitív udvariassági stratégiát képviselnek, azaz a beszélgetőpartnerek közti szolidaritást hivatottak megerősíteni. Ugyanakkor a második leggyakrabban előforduló kategória a beszélő negatív véleményét fejezi ki, akár a beszédpartner, akár külső személy vagy tény iránt, a közvetlen figyelmeztetés vagy fenyegetés azonban igen ritkán található meg a korpuszban. Ez arra utal, hogy valószínűleg a partnerek inkább egy harmadik személy vagy külső tényezők iránti nemtetszésüket fejezik ki gyakrabban.

Megemlítjük azt is, hogy az irónia és antiirónia eszközeivel viszonylag ritkán élnek élőbeszédben a beszélők, legalábbis a korpusz adatai szerint. Ugyanakkor az

Sorszám	Kategória	Gyakoriság
1.	elfogadás / egyetértés	10 659
2.	panasz / vád / kritika / sértés	3845
3.	kérés / parancs / kívánság	1875
4.	ígéret / ajánlat	1832
5.	figyelem felhívása	1566
6.	visszautasítás / egyet nem értés	1493
7.	dicséret / bók	1442
8.	üdvözlés / elköszönés	1161
9.	köszönetnyilvánítás	798
10.	bocsánatkérés	766
11.	irónia	493
12.	hárítás	299
13.	figyelmeztetés / fenyegetés	210
14.	antiirónia	24
Összesen		26 463

1. táblázat. Annotált pragmatikai egységek gyakorisága.

interaktív elemek (üdvözlés / elköszönés, valamint a figyelem felhívása) gyakori élőbeszédi sajátságnak mondhatók.

A pragmatikai szereppel bíró kifejezések leggyakoribb szavairól statisztikát is készítettünk, melyet az 1. ábra mutat be szófelhő formájában. Ebből kiviláglik, hogy elsődlegesen a köszönés, egyetértés, hezitáció és figyelemfelhívás szavai fordulnak elő. Ez részben összhangban áll a leggyakoribb annotált kategóriákkal, ugyanakkor arra is rávilágít, hogy feltehetőleg e beszédaktusoknak a leginkább korlátozott a szókincse, hiszen míg megkérni valakit vagy panaszkodni valamire sokféle nyelvi kifejezéssel lehetséges, addig például az üdvözlés vagy bocsánatkérés beszédaktusára csak viszonylag limitált számú szó és kifejezés létezik nyelvenként. Feltűnő még a diskurzusjelölők nagy száma is az ábrán, többek között a *hát*, *oké*, *egyébként* kifejezések is sűrűn szerepelnek a beszélt nyelvben.

4.2. Bizonytalansági annotáció

Kategória	Gyakoriság
weasel	7303
hedge	7166
feltételes	4885
doxasztikus	4131
peacock	2625
episztemikus	2209
kutatási	21
Összesen	28 340

2. táblázat. Bizonytalansági kategóriák gyakorisága.

1. ábra: A leggyakrabban használt szavak a pragmatikai kifejezésekben.(A „hz” a hezitálás, a „bs” pedig a beszédszándék jelölésére szolgál az annotációban.)

A 2. táblázat szemlélteti a bizonytalansági kategóriák gyakoriságát. Mindenképpen említésre méltó, hogy a rangsor első két helyét a diskurzusszintű bizonytalanság két eleme, nevezetesen a weasel és hedge kategóriák foglalják el, azaz inkább az élő diskurzusra jellemzők ezek a kategóriák. A feltételes kifejezések is gyakoriak a korpuszban, emellett a hiedelmekre utaló kifejezések is szép számmal fordulnak elő. Valószínűleg ez annak köszönhető, hogy a beszélgetések során a partnerek sokszor fejezik ki, hogy saját véleményükről, elképzelésükről van szó, nem általánosságban beszélnek. A kutatási kategória – egyáltalán nem meglepő módon – szinte alig fordul elő a korpuszban.

A 2. ábrán láthatjuk a leggyakrabban használt bizonytalansági kifejezéseket. A doxasztikus kategóriában legkiemelkedőbb talán a *szerintem* és *gondolom* szavak szerepe, amelyek a beszélő véleményét fejezik ki. A *tudom* szó egyrészt előfordulhat episztemikus kifejezésekben *nem tudom*, másrészt doxasztikus kifejezésekben is *úgy tudom*, nem meglepő módon gyakran fordul elő a korpuszban. Gyakori weasel szónak számít az *izé*, ami valószínűleg élőbeszédi sajátság, ilyen még a *nemtom* és a *tök* szó, melyek viszonylag ritkák más, bizonytalanságra annotált magyar korpuszokban.

A *kéne*, *lehetne*, *érted*, *tudod*, *hogyha*, *kicsit*, *annyira* stb. szavak mind a pragmatikai, mind a bizonytalansági szófelhőben előfordulnak, ami arra utalhat, hogy a pragmatikai és udvariassági kifejezésekben sűrűn használunk bizonytalansági kifejezéseket, ezzel enyhítve bizonyos beszédaktusok homlokzatfenyegető hatását a partnerre nézve.

2. ábra: A leggyakrabban használt szavak a bizonytalan kifejezésekben.

4.3. A bizonytalanság és pragmatikai annotáció kapcsolata

A bizonytalansági és pragmatikai annotáció összefüggéseit részletesebben is megvizsgáltuk: kigyűjtöttük a korpuszból azon bizonytalan kifejezések számát, amelyek egy pragmatikai szinten is annotált egység részeként fordultak elő (például a *Tudnál segíteni?* kérésben szerepel egy feltételes ige). A részletek a 3. táblázatban láthatók.

Az adatokból kiviláglik néhány érdekesség. Először, a feltételes bizonytalanság leginkább a kérés / parancs / kívánság, illetve az ígéret / ajánlat beszédaktusokban fordul elő a leggyakrabban, valószínűleg a gyakran használt feltételes igealakoknak köszönhetően. Másodszor, a doxasztikus (a beszélő hiedelmeire utaló) bizonytalanság leggyakrabban a bocsánatkérés, visszautasítás és a hárítás beszédaktusokban tűnik fel, amit a *hisz* ige gyakori jelenléte magyarázhat: az *azt hittem/hiszem, hogy...* frázis sokszor enyhíti a homlokzatfenyegető aktus élét. Harmadszor, megjegyezzük, hogy a peacock kifejezések leginkább az ironikus, illetve antiironikus kifejezésekben találhatók a legnagyobb arányban: az (anti)ironikus megnyilatkozásokban sűrűn szerepelnek túlzások, így peacock kifejezések is a használt kifejezések és a mögöttes tartalom közti ellentét hangsúlyozására. Negyedszer, a hedge és weasel kategóriák esetében nem láthatunk különösebben eltérő használati gyakoriságot a pragmatikai kategóriák függvényében, viszont e kettő a leggyakoribb bizonytalansági kategória szinte mindegyik pragmatikai kategóriára nézve.

Kategória	**hedge**	**weasel**	**peacock**	**episzt.**	**doxaszt.**	**felt.**	**kut.**	**össz**
elfogadás/egyetértés	107	62	97	58	93	67	0	484
antiirónia	0	1	3	0	0	0	0	4
bocsánatkérés	37	34	16	12	73	24	0	196
figyelem felhívása	1	6	0	2	11	2	0	22
hárítás	24	9	0	9	42	6	0	90
panasz/vád/kritika/sértés	1199	987	689	184	526	539	1	4125
dicséret/bók	382	240	244	46	121	145	0	1178
üdvözlés/elköszönés	7	1	3	1	9	5	0	26
irónia	81	44	96	11	13	29	0	274
ígéret/ajánlat	233	121	25	61	37	171	0	648
visszautasítás/egyet nem értés	76	64	31	41	77	39	0	328
kérés/parancs/kívánság	114	101	26	21	48	219	0	529
köszönetnyilvánítás	12	0	2	0	0	0	0	14
figyelmeztetés/fenyegetés	32	21	18	11	3	25	0	110
Összesen	2305	1691	1250	457	1053	1271	1	8028

3. táblázat. A bizonytalanság előfordulása pragmatikai kifejezésekben.

5. Összegzés

A dolgozatban bizonyos pragmatikai és szemantikai sajátságokat vizsgáltunk magyar nyelvű, nagy méretű spontánbeszéd-korpusz (StaffTalk) alapján. A vizsgálati korpusz, amely egyedülálló a magyar nyelvű spontánbeszéd-adatbázisok körében, számos szemantikai és pragmatikai sajátság kézi annotációcióját tartalmazza. A korpuszt ezek alapján az annotációk alapján, elsősorban kvantitatív szempontból elemeztük a jelen dolgozatban. Bemutattuk a leggyakrabban használt pragmatikai jelenségeket, valamint a nyelvi bizonytalanság néhány élőbeszédi érdekességére is felhívtuk a figyelmet.

A jelen dolgozatban nem volt mód arra, hogy az egyes kategóriák elemeit kvalitatív és kvantitatív szempontból, alaposabban, a szakirodalmi megállapításokkal behatóan összevetve elemezzük. Tekintettel arra, hogy a korpusz annotációjának részletessége és a korpusz méretei nemzetközi szinten is kiemelkedőek, a kutatás következő lépéseként ezeket a vizsgálatokat tervezzük elvégezni. Ahogyan arra az eredmények tárgyalásában is igyekeztünk utalni, mindezek az elemzések számos fontos adalékot adhatnak a szemantikai és pragmatikai kutatásokhoz a jövőben.

Köszönetnyilvánítás

A korpusz létrehozását az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programja 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 megköszönni a korpusz annotátorainak kitartó és áldozatos munkáját.

Hivatkozások

Andersson, L.M., Pearson, C.M.: Tit for tat? the spiraling effect of incivility in the workplace. Academy of management review 24(3), 452–471 (1999)

Austin, J.L.: How to do things with words, vol. 88. Oxford university press (1975)

Brown, P., Levinson, S.C.: Universals in language usage: Politeness phenomena. In: Questions and politeness: Strategies in social interaction, pp. 56–311. Cambridge University Press (1978)

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)

Cheng, W., Greaves, C., Warren, M.: The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). ICAME journal 29, 47–68 (2005)

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

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., Potts, C.: A computational approach to politeness with application to social factors. arXiv preprint arXiv:1306.6078 (2013)

De Felice, R., Darby, J., Fisher, A., Peplow, D.: A classification scheme for annotating speech acts in a business email corpus. Icame Journal 37, 71–105 (2013)

Deme, A., Markó, A.: Lengthenings and filled pauses in Hungarian adults' and children's speech. KTH Royal Institute of Technology (2013)

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)

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

Holmes, J., Stubbe, M.: Power and politeness in the workplace: A sociolinguistic analysis of talk at work. Routledge (2015)

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)
Leech, G., McEnery, T., Weisser, M.: Spaac speech-act annotation scheme. University of Lancaster (2003)
Leech, G.N.: Principles of pragmatics. Routledge (2016)
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)
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)
Paul, G.H., és mtsai: Logic and conversation. Syntax and semantics 3, 41–58 (1975)
Reichel, U.D., Mády, K.: Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech (2013), http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-18043-4
Rogers, P.S., Lee-Wong, S.M.: Reconceptualizing politeness to accommodate dynamic tensions in subordinate-to-superior reporting. Journal of Business and Technical Communication 17(4), 379–412 (2003)
Roque, A., Tsuetaki, A., Sarathy, V., Scheutz, M.: Developing a corpus of indirect speech act schemas. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 220–228 (2020)
Searle, J.R.: Indirect speech acts. In: Speech acts, pp. 59–82. Brill (1975)
Szabó, M.K., Vincze, V., Ring, O., Üveges, I., Vit, E., Samu, F., Gulyás, A., Galántai, J., Szvetelszky, Zs., Bodor-Eranus, E.H., Takács, K.: StaffTalk: magyar nyelvű spontán beszélgetések korpusza. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)
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)
Vincze, V.: Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013. pp.

383–391. Asian Federation of Natural Language Processing / ACL (2013), https://www.aclweb.org/anthology/I13-1044/

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

Wolfson, N.: The bulge: A theory of speech behaviour and social distance. In: Fine, J. (szerk.) Second Language Discourse: A Textbook of Current Research, pp. 21–38. Ablex, Norwood, N.J. (1988)

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)