A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata

Vincze Veronika[1], Üveges István[2,3], Szabó Martina Katalin[3,4], Takács Károly[4,5]

[1]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.
[2]Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
6722 Szeged, Egyetem utca 2.
[3]Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.
[4]Társadalomtudományi Kutatóközpont, CSS-RECENS
1097 Budapest, Tóth Kálmán utca 4.
[5]Linköpingi Egyetem, The Institute for Analytical Sociology
601 74 Norrköping, Svédország
{vinczev,martina}@inf.u-szeged.hu
uvegesistvan898@gmail.com
Szabo.Martina@tk.hu
karoly.takacs@liu.se

Kivonat A tanulmányban egy nagyméretű, magyar, beszélt nyelvi adatbázist elemzünk, és annak morfológiai és szófaji sajátságait vetjük össze más írott nyelvi korpuszok sajátságaival. A HuTongue korpusz, amelyet manuálisan leiratoztattunk és annotáltattunk, elsősorban abból a célból készült, hogy egy alapvetően szociológusokból álló kutatócsoport a pletykadiskurzusok sajátosságait vizsgálhassa (Galántai és mtsai, 2018). A korpusz szövegei hétköznapi szituációkban, külső ingerektől elszigetelt környezetben keletkeztek (Gulyás és mtsai, 2018). Legjobb tudomásunk szerint a HuTongue az első olyan, nagyméretű, magyar beszélt nyelvi korpusz, amely szűretlenül tartalmazza az összes, az adott időszakban elhangzó rögzített beszélgetés részletesen annotált leiratát. Mivel az egyes beszélgetések esetenként több résztvevő oldalán is rögzítésre kerültek, így a duplikátumok kezelése a feldolgozás egy fontos, nem triviális lépése volt. A jelen részletesen tárgyalja e munkafázis megfontolásait és módszereit. Ezt követően bemutatja a létrejövő korpusz statisztikai, köztük morfológiai és szófaji alapadatait, összevetve néhány más írott korpusz alapvető adataival. Azt reméljük, hogy a korpuszunk hatékonyan támogatja majd számos különféle szociológiai és nyelvészeti probléma korpuszalapú kutatását a jövőben.

Kulcsszavak: kézzel annotált korpusz, nyelvi erőforrás, írott és beszélt nyelv, pletyka, magyar, NLP

1. Bevezetés

Manapság egyre több olyan kutatási terület van, közöttük a szociológia vagy a nyelvészet, amely mindinkább adatközpontúvá válik. Ezek a korpuszalapú és statisztikai megközelítések azonban megbízható és nagyméretű nyelvi adatbázisok létrehozását teszik szükségessé (Neuberger és mtsai, 2014). Közülük egyre több törekszik a beszélt nyelv reprezentálására (Crowdy, 1993; Hemphill és mtsai, 1990; Maekawa és mtsai, 2000; Oostdijk, 2000; Mengusoglu és Deroo, 2001; Seppänen és mtsai, 2003; Van Bael és mtsai, 2007).

A tanulmányban bemutatunk egy nagyméretű, magyar beszélt nyelvi adatbázist, amelyet manuálisan leiratoztattunk és annotáltattunk. Az adatbázis kifejezetten a pletykadiskurzusok sajátosságainak vizsgálatához készült, azonban a korpusz mérete és a szövegek sajátságai miatt számos egyéb kutatási kérdés tárgyalásához is alapot teremthet a jövőben.

A korpusz létrehozásának fő célja a pletyka fogalmi körébe sorolható megnyilatkozások egzakt vizsgálata volt, elkészítése pedig három fő szakaszra tagolódott. A feldolgozást egy előkészítési fázis előzte meg, amelyet követően a fájlokat legépelte és annotálta egy feldolgozócsoport. E munka során tehát az annotátorok nem csupán legépelték a hanganyagok verbális tartalmát, hanem kódolták a nem verbális hanghatásokat, valamint a pletykadiskurzusokat és az utóbbiak célszemélyeit is. Az utolsó fő lépésként a kutatócsoportnak ki kellett szűrnie azokat a duplikátumokat, amelyek a felvételrögzítési sajátságok miatt kerültek a korpuszba.

A jelen dolgozat kettős céllal bír: Egyrészt a cikk részletesen tárgyalja az utolsó munkafázis megfontolásait és módszereit. Másrészt bemutatja a létrejövő korpusz statisztikai, köztük morfológiai és szófaji alapadatait, összevetve néhány más írott korpusz alapvető adataival. Célunk, hogy felmérjük, milyen jellegzetes eltéréseket tapasztalhatunk a szófaji eloszlás és morfológiai jellegzetességek tekintetében az írott és a beszélt nyelv között. Azt reméljük, hogy a korpuszunk hatékonyan támogatja majd számos különféle szociológiai és nyelvészeti probléma korpuszalapú kutatását a jövőben, valamint az írott és a beszélt nyelv összevető vizsgálataihoz is adalékot szolgáltat.

2. Kapcsolódó irodalom

A korpuszok növekvő száma ellenére még mindig viszonylagosan kevés a hangzó szövegeket reprezentálók száma, különösen azoké, amelyek gépelt leiratokat is tartalmaznak. Ez az átírási eljárás magas munkaerő- és költségigényével magyarázható. Különösen csekély a magyar beszélt nyelvű beszélt korpuszok száma, és ezek is többségükben olvasott szövegekből állnak (Gósy és mtsai, 2012). Az alábbiakban csupán egy összefoglalást adunk a magyar nyelv beszélt korpuszairól, mindezek részletesebb bemutatását l. (Szabó és mtsai, 2021).

A magyar telefonbeszéd adatbázis (MTBA) telefonon rögzített olvasott szövegeket tartalmaz. Feldolgozási módját úgy alakították ki, hogy támogathassa a beszédtechnológiai kutatásokat és fejlesztéseket (Vicsi és mtsai, 2002). A Kivi

korpusz (Kugler, 2015) különféle történetek elmeséléseiből áll, míg a Budapesti Szociolingvisztikai Interjú 250 adatközlő interjúját tartalmazza (Váradi, 2003). A HuComTech multimodális korpusz körülbelül 50 órányi video- és hangfelvételből, összesen 111 formális (szimulált állásinterjúk) és 111 informális, de irányított párbeszédből áll (Pápay és mtsai, 2011).

A fentebbiektől eltérően a spontán beszédet kívánja reprezentálni a Budapesti Egyetemi Kollégiumi Korpusz (BEKK) (Bodó és mtsai, 2017) és a BEszélt nyelvi Adatbázis (BEA) (Gósy, 2013), és ezzel összefüggésben a HuTongue szövegállománya az említettek közül a két utóbbihoz áll a legközelebb. (A kutatócsoport egy újabb, hasonló korpuszáról l. (Szabó és mtsai, 2021).) A BEKK esetében az interakciókat a résztvevők saját telefonjaikon rögzítették, ezért tulajdonképpen társalgásrészleteket tartalmaz. A BEA korpusz létrehozóinak fő célja az volt, hogy fonetikai (és nem szemantikai vagy pragmatikai) vizsgálatokat tegyen lehetővé, ezért a korpuszban alkalmazott annotációt is ennek megfelelően alakították ki. A szövegek létrejöttének körülményei, illetve feldolgozásuk módja miatt azonban a fentebb említett korpuszok csupán korlátozottan alkalmasak a magyar beszélt nyelv sajátosságainak a kutatására.

Legjobb tudomásunk szerint a HuTongue az első olyan, nagyméretű, magyar beszélt nyelvi korpusz, amely tartalmazza az összes beszélgetés részletesen annotált leiratát, ezáltal képes hatékonyan támogatni számos nyelvészeti, valamint szociológiai tárgyú kutatást, valamint, mivel egy zárt közösség adott időintervallumban elhangzott valamennyi beszélgetését tartalmazza, kiemelten alkalmas lehet kvantitatív nyelvészeti elemzésekre is (Szabó és Galántai, 2017).

3. A korpusz létrehozásának menete

Ebben a részben összefoglaló jelleggel ismertetjük a korpusz létrehozásának metódusát és eszközeit (részletesen Szabó és Galántai (2017); Gulyás és mtsai (2018); Galántai és mtsai (2018); Pápay (2019)). A folyamat fő részei a következők voltak:

- előfeldolgozás: eltávolítottuk a hosszabb csendeket és felosztottuk az anyagot kisebb egységekre,
- zajok kiszűrése Python függvénykönyvtárakkal,
- a hanganyag leírása, annotálása és a fájlok adatbázisba rendezése,
- minőségbiztosítás a teljes folyamat során,
- a duplikátumok eltávolítása,
- automatikus morfológiai elemzés a magyarlanc programcsomaggal (Zsibrita és mtsai, 2013),
- kvantitatív mérések kivitelezése a korpusz szűrt verzióján.

3.1. A korpusz anyaga és az előfeldolgozási lépések

A korpusz szövegei hétköznapi kommunikációs helyzetekben keletkeztek, egy külső ingerektől elzárt környezetben (Galántai és mtsai, 2018). A magas minőségű

hanganyagot egy szórakoztatóipari cég rögzítette. A hang rögzítését 24 órában végezték, a keletkezett korpusz pedig összesen 8 egymást követő nap felvételeit tartalmazza. A felvételek készítése során mind a nyolc önkéntes résztvevő mikroportot viselt; beszélgetéseik teljes rögzítéséhez előzetes beleegyezésüket adták.

A szövegek abban a tekintetben spontán beszélgetések, hogy azok témáit és hosszát a felvételek készítői nem határozták meg, továbbá a beszéd mennyiségét sem szabályozták. A résztvevők tehát korlátozás nélkül beszéltek annyit és arról, amennyit és amiről akartak, ugyanakkor tisztában voltak azzal a ténnyel, hogy a hangjukat folyamatosan rögzítik. Emellett néhány esetben a résztvevők cselekedeteit külső irányítással befolyásolták. Mivel ily módon a szövegek keletkezési körülményei (a résztvevők motivációi, valamint az alkalmankénti külső irányítás) befolyásolhatták a beszélői megnyilatkozásokat, a korpuszt félig vagy részlegesen spontánnak kell tekintenünk (Szabó és Szvetelszky, 2019).

A korpusz előkészítésének első lépéseként a 10 másodpercnél hosszabb csendeket automatikusan eltávolítottuk (ezáltal szegmenseket képeztünk, l. lentebb), és az anyagot 60 perces egységekre osztottuk fel. A létrehozott hangzószövegkorpusz körülbelül 500 órányi anyagot tesz ki.

3.2. Hanganyagok leiratozása és annotálása

A korpusz építésének második lépéseként 18 annotátor legépelte és annotálta a fájlokat. A feldolgozási munkát a megfelelő résztvevők kiválasztása és képzése előzte meg (részletesen l. Szabó és Galántai (2017)). A munka során az annotátorok az f4transcript szoftvert[1] használták. Mind a leiratozást, mind az annotálást ennek a segítségével végezték. Az f4 szoftvert gyakran használják szociológiai tárgyú tudományos kutatási projektekben, mivel kiváló lehetőséget ad nagyobb mennyiségű hanganyag gyors és egy időben történő leiratozására és tagelésére. Mivel projektünk célja nem az volt, hogy fonetikai elemzésekhez állítsunk elő vizsgálati anyagot, a választott szoftver megfelelő volt a számunkra.

A gépelés és annotálás színvonalának biztosítása érdekében a munka során gyakran ellenőriztük a minőséget úgy, hogy bizonyos fájlokat az összes annotátorral feldolgoztattuk, majd a kimeneteket összevetettük egymással (részletesen l. Gulyás és mtsai (2018)). A leiratok pontossága érdekében az annotátorokkal rendszeres megbeszéléseken tisztáztuk az esetleges inkonzisztenciákat, a leiratozást végzők személye pedig néhány esetben cserélődött is a minőségi elvárások tarthatósága érdekében.

Az annotátorok online kapták meg a hangfájlokat, valamint az egyes audiofájlokhoz tartozó szegmenshatárokat tartalmazó egyszerűszöveg-formátumú fájlokat. Az annotátorok tehát ezeket a fájlokat töltötték be a szoftverbe, és a szövegeket a megfelelő szegmenshatárok közé gépelték a kapott utasításoknak megfelelően.

Mivel minden résztvevő viselt mikroportot és minden mikroport anyagát feldolgoztuk, nem volt szükséges az anyagok teljes tartalmát legépelni. Az alapelv az volt, hogy leírjuk azt a beszélgetést, amelyben a mikroport viselője részt vesz.

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

Az annotátorok feladata a következő három részfeladatból állt (Gulyás és mtsai, 2018):

- a hanganyagon rögzített verbális kommunikáció legépelése,
- az anyag vizsgálata szempontjából fontos, különféle információk kódolása (időbélyegek, az adott diskurzusok résztvevői, valamint a beszélgetések során jelen levő, de meg nem szólaló résztvevők),
- különféle nem verbális hanghatások tagelése az annotálási útmutatóban előre meghatározott módon (pl. suttogás, kiabálás, nevetés, sóhaj stb.),
- a pletykadiskurzusok megjelölése a pletyka célszemélyével / személyeivel egyetemben.

Ahogyan azt az annotációs folyamat fentebb részletezett lépései is mutatják, a munka célja nem kizárólag a verbális tartalmak leírása volt, hanem bizonyos, nonverbális információk annotálása is. Ezzel összefüggésben olyan nem verbális jeleket választottunk ki az annotáláshoz, amelyeknek gyakorisága vagy együttes előfordulási jellemzői a pletyka indikátoraként szolgálhatnak (Galántai és mtsai, 2018).

A gépelőknek időbélyegek segítségével el kellett különíteniük a beszélgetések egyes megnyilatkozásait, illetve összefüggő monológjait. Ezt azt jelentette, hogy egyazon időbélyeg alá kerülhetett egyetlen megnyilatkozás vagy egy összefüggő monológ, de több résztvevő megszólalását külön időbélyegek alá kellett tenni. Az adott megszólaló nevét is jelölték. Az időbélyegek nagyon fontosak voltak a számunkra, mivel ezek a címkék nyújtanak lehetőséget a különböző mikroportokon rögzített, ezáltal különálló anyagokon létező szegmensek egymáshoz illesztésére. Mindemellett, ezek a tagek tartják meg a kapcsolatot az audiofájlok és azok írott változatai között.

Azoknak a megnyilatkozásait, akik nem voltak a vizsgálati csoport tagjai, egy speciális annotációs címkével látták el. Emellett azokat a személyeket is annotálták, akik nem szólaltak meg egy adott beszélgetés során, azonban jelen voltak (taggel jelölték a nevüket, vagy ha nem voltak azonosíthatóak, legalább a becsült számukat).

Azt is annotálni kellett, ha egy megnyilatkozás vagy egy beszélgetés egésze vagy egy része érthetetlen volt. Ezen túlmenően, ha az annotátor nem volt biztos abban, hogy jól értette az elhangzottakat, bizonytalanságát egy speciális nyitó- és zárótaggel jelölte. A nem verbális hangok (pl. köhögés, nevetés) két alapvető típusát különbözőképpen kódoltattuk: a pillanatnyit és a hosszabb ideig tartót.

Az annotációs folyamat kardinális lépése volt azoknak a megszólalásoknak a címkézése, amelyben a résztvevők valamely más, jelen nem levő résztvevőre utalnak. Ebben a vizsgálatban elsődlegesen a csoporton belüli pletykára fókuszálunk, így Kurland és Pelled (2000) alapján a pletykát kutatócsoportunk a következőképpen határozta meg: megnyilatkozás vagy beszélgetés valamely csoport általában néhány tagja között az adott csoport más olyan tagjáról vagy tagjairól, aki vagy akik nincs(enek) jelen. Amennyiben a pletyka célszemélye az annotátor számára egyértelmű volt, akkor ezt egy megfelelő annotációs címkével ugyancsak fel kellett tüntetnie.

Ahogyan az a munkafolyamat vázlatából is kitűnik, az annotálás a leiratozással, így a felvett anyag hallgatásával egy időben zajlott, tehát nem utólag végeztettük a munkát a gépelt anyagon. Az annotátoroktól azt kértük, hogy az annotálás során a hangsúlyt és a hanglejtést éppúgy vegyék figyelembe, és azok segítségével próbálják megérteni a szó szerinti jelentésen túli, szándékolt tartalmakat is, valamint azokra támaszkodva hozzanak döntést a kétes esetekben.

3.3. A duplikátumok eltávolítása

Mivel minden résztvevő mikroportot viselt, néhány beszélgetést többször is rögzítettek a felvételeken. A feldolgozást nehezítette az a körülmény, hogy ezen rögzített „beszélgetéspéldányok" száma nem egyezett meg az abban részt vevők tényleges számával. Az eltérés számos faktorból adódhatott, mint például:

- az egyik résztvevő felvételén bizonyos részletek túlságosan halkak voltak / nem álltak rendelkezésre a pontos leirat elkészítéséhez,
- túlságosan erős háttérzaj (ilyen esetekben egyáltalán nem készülhetett leirat),
- esetenként leiratozói hanyagság miatt.

A felvételek résztvevőnként eltérő hossza és szegmentálása miatt nem volt egyértelmű továbbá, ha egy adott beszélgetést már korábban rögzítettek valahol a korpuszban. Az annotátorok ezért azt az utasítást kapták, hogy írjanak le minden elhangzottat, függetlenül attól, hogy az adott beszélgetést már esetlegesen hallották egy másik mikroporton keletkezett anyag leiratozásakor, ugyanakkor feltéve, hogy az adott beszélgetésben a mikroport viselője megítélésük szerint részt vesz, valamint a beszélgetés felismerhető minőségben szerepel az adott felvételen (pl. nem túl távoli vagy zajos).

Ezeknek a duplikátumoknak a kiszűrése nyilvánvalóan kardinális feladat, hiszen enélkül a kvantitatív eredmények bármely kutatási kérdés vonatkozásában szignifikánsan eltérhetnek az ismétlődésmentes változat eredményeitől. Annak céljából tehát, hogy a vizsgálataink előtt a korpuszból a duplikátumokat eltávolíthassuk, a következő eljárást alkalmaztuk.[2] A fájlokat a 8 napon történő rögzítés okán 8 csoportba soroltuk: mindegyik csoport az adott napon rögzített felvételekből állt. Ezután összegyűjtöttük az ugyanazon a napon rögzített összes beszélgetés szókincsét a szegmenshatárokkal egymástól elválasztott diskurzusonként, azaz minden beszélgetéshez készítettünk egy szógyakorisági listát (bag-of-words). Ezután összehasonlítottuk az egyes beszélgetések szókincsét az összes többi beszélgetés szókincsével, amelyeket a többi résztvevő mikroportjai rögzítettek ugyanazon a napon. Ha két diskurzus között a szókincs legalább 75%-a egyezett, és az adott beszélgetések legalább 10 szót tartalmaztak, akkor az adott két beszélgetést azonosnak tekintettük[3]. Az esetek többségében egy hosszabb beszélgetés tartalmazott egy rövidebb szekvenciát, azaz a rövidebb szekvencia

[2] Az volt a célunk, hogy minél egyszerűbb és hatékonyabb megoldást válasszunk.
[3] Több küszöbértékkel is kísérleteztünk, és 75% bizonyult a leghatékonyabbnak.

megismétlődött a korpuszban. Ezeknek az ismétléseknek az eltávolítása érdekében a rövidebb beszélgetést töröltük az adatokból.[4]

A fenti megközelítés hatékonyságát manuálisan kiértékeltük az adatok egy kis részhalmazán; az ellenőrzéshez 50 diskurzuspárt ellenőriztünk manuálisan. Megállapítottuk, hogy 45 esetben (90%) a törölt beszélgetést egy másik, hosszabb beszélgetés valóban tartalmazta. Azt mondhatjuk tehát, hogy megközelítésünk képes volt 90%-os pontossággal (precision) megtisztítani a korpuszt az eredményeket torzító duplikátumoktól.

Módszerünkkel a három vagy annál többször előforduló szövegrészeket is lehetséges volt eltávolítanunk. A páronkénti összehasonlítás során ugyanis értelemszerűen minden diskurzust minden diskurzussal összevetettünk, így többszörös ismétlődések esetében is csupán a meghatározottak szerint legmegfelelőbb példányt tartottuk meg.

A munka során a kiinduló adatbázis 35,7%-át töröltük (15470-ből 5519 szegmens), így a HuTongue fennmaradó része (1 469 558 token) már alkalmasabb lehet más, már létező korpuszokkal való összevetésre.

A duplikátumok eltávolításának a bemutatott megoldáson túl több alternatívája is lehetséges. Az egyik lehetőség az, ha az annotációkat súlyozzuk a diskurzusban résztvevők számával. Ez azt jelenti, hogy amennyiben egy beszélgetésben például öten vettek részt, úgy az adott beszélgetés leiratában annotált információkat 1/5 részben számítjuk bele a statisztikai adatokba. A lehetséges további alternatívákkal, azok alkalmazhatóságával azonban e dolgozat keretei között nem foglalkozunk.

4. A szűrt korpusz alapvető statisztikai adatai

Ebben a fejezetben a HuTongue szűrt változatának alapvető statisztikai adatait mutatjuk be.

4.1. Annotált elemek

A korpusz annotálása során összesen 78 486 taget helyeztünk el. A nem verbális hangok statisztikai alapadatait az 1. táblázat tartalmazza.

Látható, hogy összességében több mint 50 000 hangeffektus található a rögzített anyagban, azaz a spontán beszéd számos ponton tartalmaz nem verbális elemeket. Ezek egy része a szórakoztatóipari műsor jellegéből adódik (pl. ujjongás, fütyülés, sikítás), míg más hangeffektusok előfordulhatnak nem spontán jellegű rögzített beszédben is (pl. köhögés egy hírműsor felvételében). Az automatikus beszédfelismeréshez azonban ezen elemekre mindenképpen érdemes figyelmet fordítani.

A pletykaannotáció statisztikai alapadatait a 2. táblázat tartalmazza. A neveket adatvédelmi okok miatt lecseréltük.

[4] Mivel az időbélyegek relatív, és nem abszolút időhatárokat jelöltek, nem volt lehetőség egyszerű időbélyegalapú szűrésre.

Típus	**Előfordulás**
köhögés	936
sóhajtás	3333
nevetés	32777
sírás	505
gunyoros nevetés	1420
zavarodott nevetés	1870
sikítás	579
ásítás	293
pisszegés	28
ujjongás	875
torokköszörülés	2118
fütyülés	596
éneklés	6441
összesen	51771

1. táblázat. Nem verbális hangok eloszlása.

Az adatokból egyértelműen látszik, hogy bizonyos személyeket (Zoli, Maja, András) jelentősen többször említenek, mint másokat – nevezetesen, a tagek közel 48%-a róluk szólt. Valószínűleg ők állnak a társaság életének középpontjában, több és erősebb kapcsolati hálóval rendelkeznek, mint a kevésbé gyakran emlegetett személyek. Ezzel szemben Zsani, Zsáklin, Viola és Dóri a csoport marginálisabb tagjának számítanak, ők kevésbé képezték a dialógusok központi témáját. A csoportközi említéshálózat mélyebb elemzésével egy másik dolgozatban foglalkozunk (Üveges és mtsai, 2021).

Célszemély	**Előfordulás**
Zoli	4430
Maja	3775
András	3502
Gabi	2406
Dani	1930
Vanda	1760
Kornél	1689
Sanyi	1199
Tomi	1157
Erika	925
Levi	605
Zsani	472
Zsáklin	421
Viola	219
Dóri	160
összesen	24650

2. táblázat. A pletykaszövegek eloszlása célszemély szerint.

4.2. Szófaji eloszlás

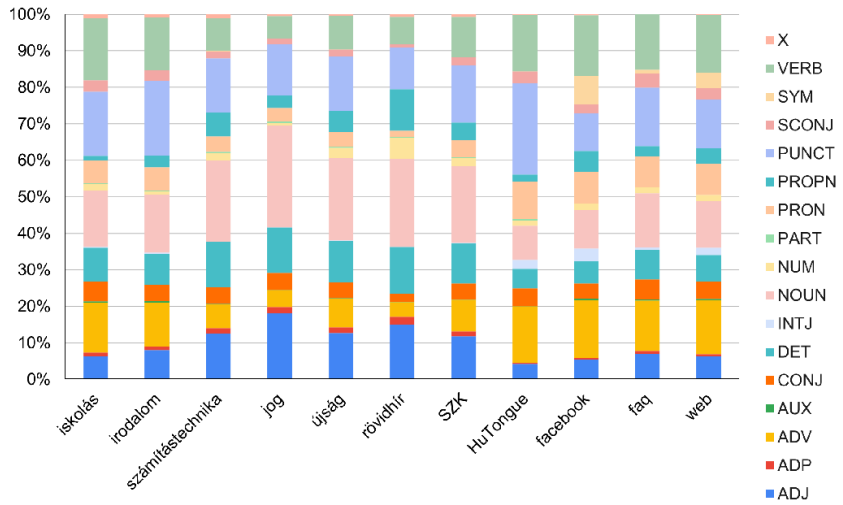
A HuTongue korpusz szövegeit a magyarlanc nyelvi elemzővel (Zsibrita és mtsai, 2013) elemeztük, majd az így kapott szófaji eloszlást összevetettük több kézzel annotált írásos szöveg szófaji eloszlásával. Ezzel azokra a kérdésekre kerestük a választ, hogy milyen jellegzetes eltéréseket tapasztalhatunk szófaji eloszlás terén az írott és beszélt nyelv között. Vizsgálatainkhoz a HuTongue korpusz mellett a Szeged Korpusz univerzális morfológiára (Vincze és mtsai, 2017) annotált változatára, valamint két kisebb, közösségi médiából származó adatbázisra (Vincze és mtsai, 2014) támaszkodtunk. E két utóbbi korpusz a Facebookról gyűjtött nyilvános posztokat, valamint a www.gyakorikerdesek.hu oldalról származó kérdéseket és válaszokat tartalmaz, és szintén az univerzális morfológia szerint lettek kézzel annotálva.

A fent említett korpuszok szófaji statisztikáit a 3. táblázat, valamint az 1. ábra tartalmazza. A korpuszok hasonlóságát az úgynevezett Kendall-együtthatóval számszerűsítettük, lásd a 4. táblázat.

Szófaj	**iskolás**	**irodalom**	**sz.tech.**	**jog**	**újság**	**rövidhír**	**SzK**	**HuTongue**	**FB**	**FAQ**	**Web**
ADJ	21267	18641	26496	46190	27799	33698	174091	61701	467	681	1148
ADP	3304	2510	2934	4269	3597	4936	21550	4411	38	78	116
ADV	46592	28201	14099	12006	17275	9093	127266	227503	1369	1332	2701
AUX	797	807	153	57	225	79	2118	562	29	33	62
CONJ	18816	10587	9551	11856	9553	5141	65504	71403	363	532	895
DET	31253	19793	26160	31495	25196	29027	162924	79276	530	785	1315
INTJ	738	814	114	6	135	5	1812	36514	288	60	348
NOUN	52385	37299	47041	71445	49590	54445	312205	136819	921	1441	2362
NUM	6102	2040	4125	1695	6246	13128	33336	21077	151	152	303
PART	956	884	709	1077	642	505	4773	7108	0	0	0
PRON	21227	14654	9188	9585	9001	3646	67301	149479	746	834	1580
PROPN	3901	7702	13807	8638	12553	25861	72462	27778	487	271	758
PUNCT	59420	47990	31241	35820	32902	25755	233128	367820	904	1551	2455
SCONJ	10521	6761	3978	4001	4142	1994	31397	48155	202	366	568
SYM	0	0	350	1	1	59	411	33	670	102	772
VERB	57905	33998	18805	15500	20526	16834	163568	227023	1440	1477	2917
X	3496	1930	2222	1386	794	1633	11461	2896	19	0	19

3. táblázat. A HuTongue, a Szeged Korpusz egyes doménjeinek, valamint egyes közösségimédia-szövegek szófaji megoszlása.

Az eredmények azt mutatják, hogy több szembetűnő szófaji gyakorisági különbséget is tapasztalhatunk az írott és beszélt nyelv között. A Szeged Korpusz egészét tekintve az írott nyelvben átlagosan jelentősen több melléknév és főnév fordul elő, míg a beszélt nyelvben az igék, határozószók, indulatszavak és névmások szerepe nő meg. Érdekes ugyanakkor megfigyelni, hogy a Szeged Korpusz egyes doménjei is eltérően viselkednek e téren. A Kendall-együtthatót is figyelembe véve a HuTongue szövegeihez a gyakori kérdések, az irodalmi, valamint az iskolás alkorpuszok állnak a legközelebb. A legnagyobb távolságot pedig a szófaji eloszlás terén az üzleti rövidhírek mutatják.



1. ábra: A szófajok eloszlása.

A kommunikatív célok alapján a vizsgált korpuszokat két nagyobb csoportra oszthatjuk:

- Interaktív korpuszok: a HuTongue mellett ide sorolhatjuk az iskolás, irodalmi, gyakori kérdések korpuszokat is. Elsődleges jellemzőjük, hogy a szerzőnek / beszélőnek határozott szándéka, hogy megszólítsa, illetve párbeszédet folytasson az olvasóval / beszélgetőpartnerrel. A HuTongue és a gyakori kérdések esetében ez a párbeszédes forma magától értetődik, ugyanakkor az irodalmi szövegekben (regényekben) is számos párbeszédes rész található. Az iskolás szövegek létrehozásakor a tanulók pedig azt az instrukciót kapták, hogy meséljenek a hallgatóságnak egy számukra kedves napról, illetve érveljenek egy téma mellett. Mindkét szövegfajtában számos, a közönség felé szóló „kiszólást" találunk a korpuszban. E domének hasonlóságát a Kendall-együttható is alátámasztja.
- Leíró korpuszok: a jogi szövegek, újsághírek, üzleti hírek és számítástechnikai szövegek fő célja az olvasó tényszerű informálása, azonban az interakció szerepe itt jóval kisebb, az olvasó szerepe majdnem kizárólagosan az információ befogadására korlátozódik. Meg kell említenünk ugyanakkor, hogy az újságok és a számítástechnikai magazinok interjúkat is tartalmaznak, melyek a párbeszédes forma miatt közelebb állnak az interaktív korpuszokhoz, így e két domén némileg közelebb áll az interaktív szövegekhez, ahogy a Kendall-együttható is mutatja.

A Facebookról származó szövegek ugyancsak változatosak az interakció szempontjából: egyrészt különféle márkákat, sztárokat stb. képviselő oldalak nyilvános informatív bejegyzései kerültek ide (kommentek nélkül), másrészt személyes jellegű (de nyilvános láthatóságú) bejegyzéseket is találunk itt. A kétfajta bejegyzés

célja megint csak eltérő, így nem meglepő, hogy a Facebook-szövegek is valahol középen helyezkednek el az együtthatósági skálán.

Az interaktív szövegek szófaji jellemzői tehát az alábbiakban foglalhatók össze. Gyakoriak bennük az igék (az olvasót / hallgatót cselekvésre buzdítják) és a névmások (élőbeszédben vagy az ahhoz közel álló írott szövegekben gyakoriak a deiktikus utalások). Az indulatszavak szerepe is kiemelkedő, ezek nyomatékosítják adott esetben a mondanivalót, máskor figyelemfelhívó szereppel bírnak stb. A határozószavak gyakorisága pedig az igék gyakoriságával függhet össze: igék mellett határozószavak jelennek meg, ellenben főnevek mellett melléknevek tudják kifejezni ugyanazt a minőséget. Míg a leíró jellegű korpuszokban a főnevek és melléknevek szerepe domináns, addig ugyanazt a jelentéstartalmat a dinamikusabb ige + határozószó pár fejezi ki az interaktív korpuszokban.

Korpuszrész	**Hasonlóság**
iskolás	0,9534
irodalom	0,9559
számítástechnika	0,8909
jog	0,9007
újság	0,9007
rövidhír	0,7978
Szeged Korpusz	0,9081
Facebook	0,8848
Gyakori kérdések	0,9558
Webes szövegek	0,9363

4. táblázat. A HuTongue hasonlósága a Szeged Korpusz egyes doménjeihez, valamint egyes közösségimédia-szövegekhez a szófaji eloszlás alapján.

5. Összegzés

A dolgozatban bemutattuk és kontrasztív módon elemeztük a HuTongue korpusz újabb, duplikátumoktól megtisztított változatát. A HuTongue a magyar beszélt nyelvet reprezentálja, a hangzó szövegek legépelt és annotált változatával együtt.

A tanulmány célja az volt, hogy összefoglalja a korpuszkészítés fő lépéseit és módszereit, majd ismertesse azt a megoldást, amellyel sikeresen kiszűrtük a korpuszban található többször előforduló azonos szövegrészeket. A korpusz a szöveganyaga, mérete, valamint a szövegek feldolgozási módja miatt lehetőséget ad számos olyan kutatás elvégzésére, amely a magyar beszélt nyelv valamely sajátságát veszi górcső alá. Mindemellett a duplikátumok kiszűrésével a korpuszelemzés kvantitatív eredményei is mentesülnek a torzító tényezők alól.

Második fő lépésként bemutattuk a korpusz újabb változatának alapvető adatait, és azokat összevetettük néhány más szövegkorpusz morfológiai és szófaji gyakorisági sajátságaival. Az összevetés során rámutattunk néhány olyan hasonlóságra és eltérésre, amely az egyes korpuszok, illetve szövegtípusok és -domének

között mutatkozik. Bízunk benne, hogy a HuTongue korpusz további érdekes adalékokkal járulhat hozzá a magyar spontán beszélt nyelv különböző vizsgálataihoz.

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

Köszönetnyilvánítás

A kutatást az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében támogatta (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.

Hivatkozások

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)

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

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)

Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T.E., Beke, A., Neuberger, T., Nikléczy, P.: Bea: Beszélt nyelvi adatbázis (2012)

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 (1990)

Kugler, N.: Megfigyelés és következtetés a nyelvi tevékenységben. Tinta Könyvkiadó (2015)

Kurland, N.B., Pelled, L.H.: Passing the word: Toward a model of gossip and power in the workplace. Academy of management review 25(2), 428–438 (2000)

Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous Speech Corpus of Japanese. In: LREC. pp. 947–9520. Citeseer (2000)

Mengusoglu, E., Deroo, O.: Turkish LVCSR: Database Preparation and Language Modeling for an Aglutinative Language. In: IEEE International Conference on Acoustics Speech And Signal Processing. vol. 6, pp. 4018–4018. IEEE (2001)

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, B.: The Purpose and Types of Organizational Gossip. Ph.D.-értekezés (2019)

Pápay, K., Szeghalmy, Sz., Szekrényes, I.: Hucomtech multimodal corpus annotation. Argumentum 7, 330–347 (2011)

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)

Szabó, M.K., Szvetelszky, Zs.: Részlegesen spontán körülmények között keletkezett pletykaszövegek pragmatikai szempontú vizsgálata. Nyelvtudományi Közlemények 115, 317–343 (2019)

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)

Üveges, I., Szabó, M.K., Vincze, V.: Szó, beszéd – avagy hogyan kommunikálunk egymásról. 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)

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., Simkó, K.I., Szántó, Zs., Farkas, R.: Universal Dependencies and morphology for Hungarian - and on the price of universality. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 356–365. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-1034

Vincze, V., Simkó, K.I., Varga, V.: Annotating uncertainty in Hungarian web-text. In: Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop. pp. 64–69. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (Aug 2014), https://www.aclweb.org/anthology/W14-4909

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. pp. 763–771 (2013)