Felszólításannotálás a MedCollect egészségügyi álhírkorpuszban[1]

Szécsényi Tibor, Nagy C. Katalin, Németh T. Enikő

SZTE Általános Nyelvészeti Tanszék
MTA-SZTE-DE Elméleti Nyelvészeti és Informatikai Kutatócsoport
{szecsenyi, nemethen}@hung.u-szeged.hu
nagykati@hist.u-szeged.hu

Kivonat: Napjaink egyik legégetőbb problémája az álhírek terjedése, és ezáltal a hírolvasó emberek félrevezetése. Az álhírek elleni küzdelem egyik lehetséges eszköze a megjelenő hírek tényellenőrzésén kívül az álhírek automatikus felismerése. Ez történhet az álhírek szókészletét figyelembe vevő szövegosztályozókkal is, de ezek pontosságát javíthatja az álhírekben megfigyelhető nyelvhasználati stratégiák és az azokat megvalósító nyelvi elemek azonosítása. A tanulmányban bemutatjuk az ilyen stratégiák és elemek azonosítását célul kitűző magyar nyelvű MedCollect egészségügyi álhírkorpusz építését, és benne a felszólítást végrehajtó nyelvi elemek funkcióinak a kézi annotálását. A korpusz az álhírekre jellemző nyelvhasználati stratégiák és nyelvi eszközök kvalitatív és kvantitatív vizsgálatát teszi lehetővé.

1 Bevezetés

Korunk egyik legnagyobb kihívása a digitálisan terjesztett információk világában való eligazodás. A digitális platformokon a valódi híreknél mintegy tízszer nagyobb eléréssel terjednek az álhírek és az áltudományos szövegek (Krekó és mtsai., 2021.04.15), gyakran nagyon súlyos károkat okozva mind az egyén, mind a társadalom számára az élet különböző területein (egészségügy, demokrácia, nemek közötti egyenlőség, klímaváltozás elleni fellépés, biztonságpolitika stb.), veszélyeztetve a fenntartható fejlődési célokat (United Nations, 2023). A dezinformációnak elsősorban az a több mint 5 milliárd ember van kitéve, aki használja az internetet, közülük különösen érintett az a 4,7 milliárd, aki a közösségi médiát is (Soltész, 2023). A COVID–19 járvány alatt megmutatkozott, hogy milyen súlyos károkat okoznak az álhírek az egészségügy területén. 2020 első 3 hónapjában a világon mintegy 6000 fő került kórházba és 800-an meghaltak az álhírek miatt (Islam és mtsai., 2020). Az orosz-ukrán háborúhoz kapcsolódóan is hatalmas mennyiségű álhír zúdul a világra, aminek a következményei beláthatatlanok (Huszák, 2022). Az online dezinformáció által okozott károk nagyságát felismerve az Európai Unió Tanácsa 2015-től kezdve folyamatosan küzd a dezinformálás és annak következményei ellen cselekvési tervek, riasztási rendszerek, tényellenőrző szolgáltatások stb. létrehozásával (Soltész, 2023). Az ENSZ pedig jelenleg készíti a tagállamainak szóló ajánlást a dezinformáció elleni védekezésre vonatkozóan (United Nations, 2023).

[1] Jelen tanulmány az MTA Tudomány a Magyar Nyelvért Nemzeti Program *Álhírek, áltudományos nézetek nyelvészeti azonosítása* című alprogramjának támogatásával készült.

Felmerül a kérdés, hogyan azonosíthatók a dezinformáció megjelenési formái, az álhírek és az áltudományos szövegek? Az MTA-SZTE-DE Elméleti Nyelvészeti és Informatikai Kutatócsoportja e kérdés megválaszolását tűzte ki célul az MTA Tudomány a Magyar Nyelvért Nemzeti Program *Álhírek, áltudományos nézetek nyelvészeti azonosítása* című alprogramjának (2022–2026) kidolgozása során. A kiinduló hipotézisünk az, hogy az álhírek és az áltudományos szövegek rendelkeznek olyan nyelvi jegyekkel és nyelvhasználati stratégiákkal, amelyek alapján vagy amelyek kombinációi alapján egy szövegről gyanítható, hogy az álhír vagy áltudományos szöveg. A kutatásunknak három fő célkitűzése van: (1) Az álhírek és az áltudományos szövegek vizsgálatának tudományelméleti megalapozása, (2) az egészségügy tématerületére tartozó álhírek és áltudományos szövegek nyelvi jegyeinek és nyelvhasználati stratégiáinak megállapítása, (3) automatikus egészségügyi álhírdetektáló eszközök fejlesztése. Az első célkitűzés megvalósítása az alapja az álhírek, áltudományos szövegek és a kontroll szövegek megkülönböztetésének (vö. Rákosi, 2023a, 2023b). A második cél megvalósításához fel kell tárnunk a szövegekben explicite azaz testes formával megjelenő nyelvi jegyeket (tipikus szókészlet, felszólításokat megvalósító morfológiai, szintaktikai és lexikai eszközök, a tegezés és magázás eszközei, szentiment kifejezések stb.) és az implicit pragmatikai jelenségeket (implicit argumentumok, közvetett beszédaktusok, implikatúrák stb.). Ezeknek a jegyeknek az együttes figyelembe vétele vezethet az álhírek nyelvi jellemzőinek feltárásához, és a harmadik célkitűzés megvalósításához. A projekt részletesebb bemutatását Németh T. (2023) nyújtja.

Jelen tanulmányunkban a fentiek közül egy nyelvi jellemzőt, a direkt felszólításokat vizsgáljuk. Előzetes kvalitatív korpuszvizsgálat alapján kialakult hipotézisünk a következő: az egészségügyi álhírek szignifikánsan több felszólító igealakkal megvalósított direktívát tartalmaznak mint a valódi hírek az olvasók felé irányuló erőteljesebb nyomásgyakorlás[2] miatt. Módszerünk korpusznyelvészeti, ezen belül a „formától a funkcióig" korpuszpragmatikai módszer (Aijmer, 2018; Németh T. és mtsai., 2020). Korpuszunk a MedCollect korpuszból (l. 2. fejezet) elkülönített részkorpusz, amely az egészségügy tématerületére tartozó álhíreket és valódi híreket tartalmaz (l. 4. fejezet). A részkorpuszt pragmatikailag annotáltuk, ami számos problémát vetett fel, mivel a felszólítások esetén nincsen egy-az-egyhez megfelelés a formák és a funkciók között (l. 3. és 4. fejezet). Végül az annotált részkorpuszt kvantitatív elemzésnek vetettük alá, ami igazolta a hipotézisünket (l. 5. fejezet).

2 A MedCollect egészségügyi álhírkorpusz bemutatása

A MedCollect egészségügyi álhírkorpuszba egészségüggyel kapcsolatos, nyilvánosan elérhető online szövegeket gyűjtöttünk. A korpusz jelenleg nem szabadon hozzáférhető (további információkért lásd: http://enyik.szte.hu). A szövegek gyűjtése során elsődleges szempont volt, hogy azok hírszerűek legyenek, a nagyközönségnek szolgáltassanak valamilyen egészségügyi témájú hírt. A korpuszba kizárólag a hírek szövegei

[2] Természetesen az erőteljes nyomásgyakorlás nem kizárólag az álhírek jellemzője lehet, és nem is minden álhírben jelenik meg. Így a felszólítások jelenléte se nem szükséges, se nem elégséges feltétele az álhírségnek, ugyanakkor egészségügyi szövegekben jó mutatója lehet annak.

kerültek bele (cím, lead, szöveg), a hírek utáni esetleges kommentek, a reklámok, a hírek mellett a weboldalon megjelenő egyéb tartalmak nem. A hírek gyűjtése 2021 májusa és 2023 áprilisa között történt. A gyűjtött hírek keletkezési dátuma ennél nagyobb időtartamot ölel fel, a legkorábbi hír 2007-es, de a szövegek 75%-a 2020 utáni.

A korpuszba álhírek és kontroll szövegekként valódi hírek kerültek, az összesen 2206 szövegből 1448 álhír, 758 kontroll szöveg. Az álhírek egyedi gyűjtésűek, egy részüket a kutatócsoport tagjai, másik részüket a kutatáshoz kapcsolódó kurzusok hallgatói találták. A kézi gyűjtésű hírek álhír-státusát az álhírek gyűjtői, a kurzus oktatói és a résztvevő kutatók közösen állapították meg. A kézzel gyűjtött hírek egy kis része a kontroll szövegek közé került, de a kontroll szövegek nagyobbik része automatikus gyűjtésű: a kutatócsoport orvos és gyógyszerész tanácsadói javaslatára néhány megbízható, egészségügyi híreket közlő portálról származnak véletlen mintavétellel (pl. pharmaonline.hu, medicalonline.hu). Az összegyűjtött szövegeken duplikátumellenőrzést végeztünk: az azonos portálról származó, .95-nél nagyobb Jaccard hasonlóságot mutató szövegekből csak egyet hagytunk a korpuszban. A MedCollect korpuszba összesen 179 különböző portálról kerültek be a szövegek, a szövegek kb. 90 százaléka 26 portálról. A cikkek gyűjtésében összesen 30 fő vett részt.

A szövegek gyűjtésénél a későbbi ellenőrizhetőség miatt a szövegekhez hozzárendeltük az URL-t, ahonnan származnak. Mivel az álhírek sokszor ideiglenesen vannak csak jelen az interneten, ezért az URL mellett PDF formátumban elmentettük a teljes weboldalt is. Későbbi vizsgálatoknál a weboldal szerkezete, a szövegek tipográfiája, az oldalon megjelenő hirdetések is hasznos információval szolgálhatnak. Bár a weboldal eredeti HTML formátumban történő elmentése további információkat is szolgáltathatott volna, a kutatás első fejezetben ismertetett célkitűzéseinek a megvalósításához elegendő volt a PDF formátum is. Mivel a teljes weboldal elmentése során az oldalon található felugró reklámok sokszor kitakarják a lényeges tartalmat, ezért az oldalnak azt a részét külön is elmentettük PDF-ként, amely a cikk szövegét tartalmazta.

A gyűjtött híreket szöveges formátumban is eltároltuk, amelyet az eredeti weboldalról nyertünk ki. A szöveges mentésnél töröltük a szövegbe ékelődő hirdetéseket, képeket, illetve a szöveg forrására utaló tartalmakat: a szöveg szerzőjét, a szöveg megosztására buzdító, a portál minden oldalán megtalálható részeket, a szövegvégi forrásmegjelölést (pl. MTI).

A MedCollect híreinél metaadatként elérhető a hír azonosítója, a letöltés ideje, a gyűjtő azonosítója, a hír címe, URL-je, a hír megjelenésének a dátuma (ha elérhető), az elérhető formátumok (txt, teljes pdf, rövid pdf), illetve az álhírstátusz (álhír, kontroll).

A korpusz szövegeinél az eredeti szöveges változaton kívül az e-magyar automatikus elemzővel (Váradi és mtsai., 2018) elemzett emtsv változatot (Indig és mtsai., 2019) is generáltunk. Ezt felhasználva a korpusz anyagából könnyen tudunk összetett mintákkal leírható példákat keresni további kvalitatív vagy kvantitatív elemzésekhez. Az emtsv változatban található adatok alapján a MedCollect korpusz összesen 1 270 080 tokenből áll, amelyből 874 985 token az álhír, 359 095 token a kontroll szöveg. Az átlagos szöveghossz 576 token, a leghosszabb szöveg 9784 tokenes.

A teljes MedCollect korpuszból kijelöltünk egy 707 szövegből (370 300 token) álló részkorpuszt kézi annotálásra, amelyből 322 szöveg (182 674 token) álhír, 385 szöveg (187 626 token) kontroll szöveg. Ezen a részkorpuszon végeztük el a felszólító alakok annotációját is.

3 A felszólításannotálás során használt címkék

A kutatócsoport munkájának kiinduló hipotézise az, hogy az álhírek és az áltudományos szövegek rendelkeznek olyan nyelvi jegyekkel és nyelvhasználati stratégiákkal, amelyek alapján vagy amelyek kombinációi alapján egy szövegről gyanítható, hogy az álhír vagy áltudományos szöveg.

Az első nyelvhasználati stratégia, amelynek elemzésére vállalkoztunk, a direktívák, azaz a felszólítások használata volt. A direktívák olyan megnyilatkozások, amelyek funkciója, hogy a hallgatót rávegyék egy jövőbeli cselekedet végrehajtására. A felszólítás direkt vagy indirekt módon, többféle nyelvi formával vagy jeggyel kivitelezhető. Ebben a tanulmányban a felszólító alakú igéket tartalmazó megnyilatkozásokat vizsgáljuk (***Viselj*** *maszkot!*).[3] Bár a felszólító funkcióval mindig a teljes megnyilatkozás rendelkezik, a felszólító módú igék használata esetén egyértelműen ez a nyelvi elem váltja ki a felszólítást, ezért a felszólító alakú igéket jellemeztük különféle jegyekkel.

A felszólító alak jelenléte egy megnyilatkozásban nem szükségszerűen vonja maga után a megnyilatkozás felszólító funkcióját, nincsen egy-az-egyhez megfeleltetés a forma és a funkció között. A célunk elsősorban azoknak az eseteknek a vizsgálata, amelyben jelen van a felszólító funkció, mégpedig a szöveg szerzője és az olvasó viszonylatában, vagyis amikor a szerző nyomást gyakorol az olvasóra az olvasás pillanatában.

A felszólító alakok (*directive*) jellemzésére használt jegyek között első a felszólítás meglétét és minőségét kategorizáló címkekészlet. Ezek a következők: *nodirectiva, szövegszervező, interakciós, meta, saját hangú* és *közvetített*.

A *nodirectiva* címkéjű felszólító alakok egyáltalán nem hajtanak végre felszólítást. Ezek többnyire kötőmódban álló igék, amelyek alakilag megegyeznek a felszólító módú igékkel. Ezek előfordulhatnak célhatározói mellékmondatban (*Leültem, hogy* ***pihenjek***), vagy bizonyos igék *hogy* kötőszavas mellékmondati vonzatában (*Hozzálátam, hogy* ***megegyem*** *a levest*), azonban nem csak mellékmondatokban fordulhatnak elő felszólító funkció nélküli felszólító alakok, hanem főmondatban is, például tanácskérésnél: *Mit* ***tegyek****?* Ezeknek a felszólító alakoknak közös jellemzője, hogy ugyanabban a kontextusban (mondatban) általában nem cserélhetők ki a kijelentő módú alakjukra, ami a valódi felszólításoknál többnyire megtehető. Azonban nem minden kötő módú ige kap ilyen címkét, a *felszólít* mellékmondatában kötő módú ige áll, de a megnyilatkozás rendelkezik felszólító erővel: *Felszólítalak, hogy* ***viselj*** *maszkot*.

Szövegszervező címkét azok a felszólító alakok kapnak, amelyeknek lehet ugyan valamilyen felszólító erejük, elsődleges feladatuk viszont a figyelemirányítás, a koherencia megteremtése, pl. ***lásd***, ***vö***, *hogy csak párat* ***említsek***.

Az *interakciós* címkéjű felszólító alakok a beszélőnek és/vagy a hallgatónak a szöveg létrehozását és megértését segítő tevékenységével vagy a beszélő és a hallgató közt lévő viszonnyal kapcsolatosak. Reflektálhatnak a közös jelentéslétrehozás folyamatára, a beszélő és/vagy a hallgató tudására, gondolkodására, belső állapotaira, érzéseire, attitűdjeire. Egyesek közülük diskurzusjelölők, de nem mindegyik. Példák: ***Gondoljanak*** *bele! Hogy őszinte* ***legyek****..., Ne* ***feledjük****, hogy...*

[3] Más nyelvi eszközökkel, indirekt módon is végrehajthatunk felszólítást, mint például a következő konvencionálisan indirekt formával (l. Searle, 1975): *Jó lenne, ha maszkot viselnél*.

Külön csoportot alkotnak azok az alakok, amelyek egy felszólítást leíró vagy idéző megnyilatkozásban szerepelnek, amelynek azonban nincs az olvasó mint potenciális címzett felé felszólító ereje. A szöveg írója beszámol ugyan egy felszólításról, az azonban nem vonatkoztatható a jelenlegi olvasóra mint címzettre. Ilyenkor a *meta* címkét osztottuk ki. A *meta*-ként címkézett megnyilatkozások másik esete, amikor a leírt vagy idézett felszólításhoz a szöveg szerzője nem csatlakozik, annak felszólító erejét nem közvetíti az olvasó(k) felé, attól elhatárolódik, pl.: *Senkit sem szabad orvosi kísérletre kényszeríteni informált beleegyezés nélkül. Sok média, politikai és nem orvosi személy azt mondja az embereknek, hogy* ***menjenek*** *el az oltásra, az biztonságos, és nem nyújtanak információt a génterápia káros hatásairól vagy veszélyeiről.*

A korábbi csoportokba nem tartozó, azaz valódi felszólító funkciójú megnyilatkozások közös jellemzője, hogy valamilyen nyomást gyakorolnak az olvasóra az olvasás időpontjában. Ez a nyomás eredhet magától a szerzőtől, aki saját hangján, közvetlenül szólítja fel olvasóit valamire, vagy pedig lehet közvetített, amikor az írás szerzője egy másik személy/személyek felszólítását közvetíti, mert azzal egyet ért. Előbbieket a *saját hangú*, míg utóbbiakat a *közvetített* címkével láttuk el.

A *saját hangú* és a *közvetített* felszólító alakokhoz további jegyeket rendeltünk hozzá, amelyek a felszólítás forrását (*source*) és címzettjét (*target*) jellemezték.

A felszólítás forrásának jellemzésekor megkülönböztettük azt a két esetet, amikor a szöveg írója az egyedüli forrás (*speaker*), illetve amikor nem (*speaker+*). Ez utóbbi esetben is feltételeztük, hogy a szöveg írója egyetért a felszólítással, támogatja azt, ezért a *közvetített* felszólítások minden esetben *speaker+* címkét kaptak. A saját hangú felszólítás is lehet azonban *speaker+* címkéjű, ha a szöveg írója egy csoport tagjaként fogalmazza meg a felszólítást, pl. ***Kérjük****, jelentkezzen még ma.*

A *target* jegyek értéke *listener*, *listener+*, illetve *inclusive* lehet. A *listener* címkét akkor kapta egy felszólító alak, ha formailag egyedül az olvasó a címzett, még ha potenciálisan több olvasó is érintett lehet – az olvasók mégis külön-külön érezhetik megszólítva magukat.

Az *inclusive* címkét azok a felszólító alakok kapták, amelyekben az olvasó és a szöveg írója egyaránt felszólított, vagyis a szöveg alkotója az olvasóval alkotott közösségét is hangsúlyozza, pl. *A szakemberek nem tanácsolják, hogy bármit is alufóliában* ***süssünk****.*

A *listener+* címke esetében egy csoport tagjaként vonatkozik a felszólítás az olvasóra: ha a csoport tagjának tartja magát, akkor felszólítva kell éreznie magát. Pl.: *Az időseket arra kérjük, hogy fokozottan* ***figyeljenek*** *oda magukra*.

Az eddig ismertetett címkéken kívül használunk még jegyenként egy-egy címkét azokra az esetekre, amikor a felszólító alakhoz több címkét is lehetett rendelni, azaz a kategóriába sorolás nem egyértelmű. A *directive* jegyeknél ez az *ambiguous* címke volt, a *source* és a *target* jegyeknél pedig a *vague* címke.

A felszólító alakok jegyeit és lehetséges címkéit láthatjuk az 1. táblázatban.

1. Táblázat: A felszólító alakok annotálása során használt jegyek és címkék

directive	source	target
nodirectiva	_	_
szövegszervező	_	_
interakciós	_	_
meta	_	_
saját hangú	speaker speaker+ vague	listener listener+ inclusive vague
közvetített		
ambiguous	_	_

4 A felszólításannotálás menete, annotátori egyetértés

Ahogy a 2. fejezetben már említettük, a felszólító alakok kézi annotálása nem a teljes MedCollect korpuszon történt, annak egy részkorpuszát különítettük el erre a célra. A kézzel annotált részkorpusz összesen 707 szövegből áll (370 300 token), ebből 322 álhír (182 674 token), 385 pedig kontroll szöveg (187 626 token).

Az annotálás során az előző fejezetben ismertetett jegyekkel jellemeztük a szövegben található felszólító alakokat, az annotálást a WebAnno alkalmazással (Eckart de Castilho és mtsai., 2016) végeztük. Az annotátorok előannotált szövegekkel dolgoztak, az előannotált fájlokban jelölve voltak az emtsv által felszólító alaknak elemzett szavak, valamint annak a tagmondatnak az eleje és a vége is, amelyben a felszólító alak szerepelt. Az annotátorok egyrészt ezen előannotált szavakat ellenőrizték, hogy valóban felszólító alakok-e, illetve az előelemzés által meg nem talált felszólító alakokat is bejelölhették. A tényleges annotálási munka során az előző fejezetben ismertetett, előre megadott címkéket rendelhették a felszólító alakokhoz.

A felszólító alakok címkézéséhez annotálási útmutató készült 2022. december és 2023. február között. Az annotálási útmutató mintafájlok annotálása során alakult ki, amelyben a kutatócsoport tagjai vettek részt.

A tényleges annotálás 2023. februártól 2023. júliusig zajlott. A MedCollect részkorpuszának minden szövegén 3 annotátor dolgozott, az annotációkat egy kurátor hitelesítette, eltérő annotációk esetén többségi döntésre alapozva. A problémás eseteket heti rendszerességgel kutatócsoporti szinten elemeztük. Az annotálási munkálatokban 22 annotátor és 2 kurátor vett részt, az annotátorok többségében a kutatócsoport hallgatói munkatársai voltak, kisebb részben a kutatáshoz kapcsolódó szeminárium hallgatói. Az annotálási folyamat körülbelül 400 munkaórát igényelt.

Az annotátori egyetértést Cohen-féle kappa egyetértési metrikával jellemeztük, a kiszámításában a python sklearn programcsomagot használtuk. Az egyetértést nem az egyes annotátorok között értékeltük, hanem az annotátorokat a kurátori döntéshez viszonyítottuk: vettük az egyes annotátorok által annotált összes fájlt, és ez viszonyítottuk a kurátorok által jóváhagyott címkékhez, így minden annotátorhoz tartoztak egyetértési értékek. Egyetértési értéket számoltunk a jegycsoportokra vonatkoztatva

összesítve (*directive*, *source*, *target*), valamint az egyes címkékre is, így minden annotátorhoz 19 egyetértési érték tartozott. Az egyetértési értékeket nem az összes elemzett szó alapján számítottuk, hanem csak a releváns szavakat vettük figyelembe: a *directive* jegyeknél azoknak a szavaknak a címkézését, amelyeket az annotátor és a kurátor is felszólító alaknak elemzett, a *source* és *target* jegyeknél pedig azokat, amelyeket mindketten saját hangúnak vagy közvetítettnek. Az annotátorok egyetértési értékeinek átlagát tekintettük összesített egyetértési értéknek (jegyenként vettük az összes annotátor átlagát).

A *directive* jegyek annotátori egyetértési adatai láthatók a 2. táblázatban.

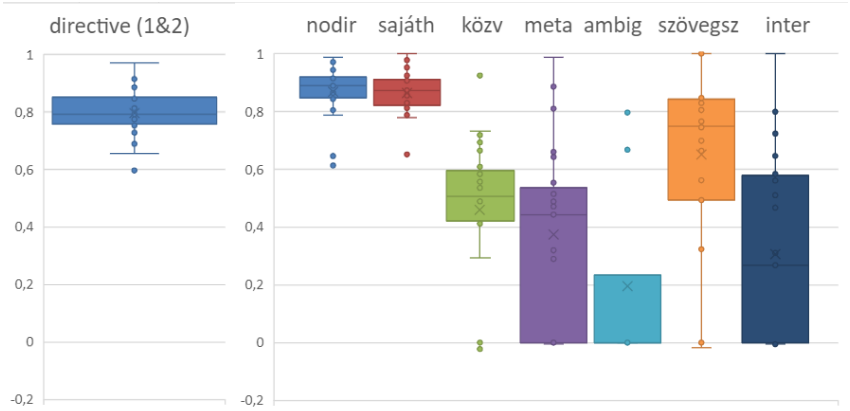
2. Táblázat: A *directive* címkékkel ellátott felszólító alakok száma és az annotátori egyetértések átlagai

directive	kurátori előfordulás	annotátori egyetértések átlaga
összesítve	2664	0,793
nodirectiva	1243	0,867
saját hangú	962	0,862
közvetített	184	0,457
meta	144	0,376
szövegszervező	70	0,645
interakciós	46	0,304
ambiguous	12	0,194

Az egyetértési értékek átlaga a *nodirectiva* és a *saját hangú* címkék esetében nagyon jónak számít (0,867 és 0,862). A *közvetített* és *meta* címkék egyetértési értékei közepesek (0,457 és 0,376), ez azt mutatja, hogy ezeket nehéz megkülönböztetni egymástól az anyanyelvi beszélőknek is. Az utolsó két címke kis előfordulási száma okozza azok csekély egyetértési értékét. Az egyetértési érték az összes értékre számolva szintén nagyon magas.

A kapott egyetértési értékek megfelelnek annak az előzetes várakozásunknak, hogy bár a felszólító alakok néhány funkciója jól elkülöníthető, más esetekben nehéz őket megkülönböztetni. Az eredmények alapján arra számítunk, hogy egy automatikus elemző a felszólító alakok közül a szavak kontextusát is figyelembe véve hatékonyan ki tudja szűrni a *nodirectiva* és a *saját hangú* címkéjűeket, esetleg még a valamilyen irányban tényleges felszólítást megfogalmazó *közvetített* és *meta* címkéjűeket is, ugyanazon csoport tagjainak tekintve őket.

Érdemes megtekinteni a *directive* jegyekkel kapcsolatos annotátori egyetértések eloszlását is (a különböző pontok a különböző annotátorok egyetértési értékeit mutatják, Excel oszlopdiagramként ábrázolva):



1. Ábra. A *directive* címkék annotálása során mért annotátori egyetértési értékek eloszlásai

Látható, hogy a *nodirectiva*, *saját hangú* és *közvetített* címkék esetében aránylag kicsi az annotátori egyetértések szórása, a többi esetben nagyobb. A nagyobb szórások, különösen a lefelé látható kiugró adatok valószínűleg az annotálási feladat rossz értelmezéséből fakadnak, illetve az annotátorok gyakorlatlanságából: a kísérő kurzus hallgatói a törzsannotátorokhoz viszonyítva kevés szöveget annotáltak.

A *source* és a *target* jegyek összesített és jegyenkénti annotátori egyetértési értékei láthatók a 3. és 4. táblázatban.

3. Táblázat: A *source* címkékkel ellátott felszólító alakok száma és az annotátori egyetértések átlagai

source	kurátori előfordulás	annotátori egyetértések átlaga
összesítve	1146	0,485
speaker	857	0,484
speaker+	288	0,492
vague	1	0

4. Táblázat: A *target* címkékkel ellátott felszólító alakok száma és az annotátori egyetértések átlagai

target	kurátori előfordulás	annotátori egyetértések átlagai
összesítve	1146	0,851
listener	573	0,94
listener+	92	0,568
inclusive	480	0,854
vague	1	0

A felszólítás forrásának jelölésénél az annotátori egyetértés közepes volt, a felszólítás megszólítottjánál viszont magas. Az annotátori egyetértések szórása a *source* jegyeknél közepes, a *target* jegyeknél a *listener+* kivételével nagyon magas. Az annotátori egyetértési értékek táblázatos formában annotátoronként és jegyenként/címkénként megtekinthetőek a https://github.com/enyik-rg/directive_annotation github repozitóriumban, az egyetértési értékek eloszlását szemléltető oszlopdiagramokkal együtt.

5 Az annotálás kvantitatív elemzése

Kiinduló hipotézisünk az volt, hogy az álhírek szignifikánsan több felszólító igealakkal megvalósított direktívát tartalmaznak, mint a valódi hírek. Ennek oka az olvasók felé irányuló erőteljesebb nyomásgyakorlás.

Az annotált korpuszban a kurátori jóváhagyást követően összesen 2664 felszólító alakot (imp) találtunk, ebből 1842 volt az álhírekben, 822 a kontroll szövegekben. Mivel az álhírek és a kontroll szövegek össztokenszáma (wnum) közel azonos, ezért a kiinduló hipotézis már ezen adatok alapján is alátámasztottnak látszik. A felszólító alakok *directive* címkéinek a számát közöljük az 5. táblázatban.

5. Táblázat: A *directive* címkével ellátott felszólító alakok száma a MedCollect korpuszban

	wnum	imp	nodir	shang	közv	meta	amb	szszerv	inter
álhír	182 674	1842	699	901	99	53	1	53	33
kontroll	187 626	822	544	61	85	91	11	17	13
mind	370 300	2664	1243	962	184	144	12	70	46

Látható, hogy míg a *nodirectiva* címkéjű felszólító alakok száma nagyságrendileg nem különbözik, a *saját hangú* felszólítások viszont mintegy tizenötször nagyobb számban fordulnak elő az álhírekben.

Mivel a két szövegtípusban eltérő a felszólító alakok (imp) száma, árnyaltabb képet kapunk, ha az egyes címkék ezekhez viszonyított számát vizsgáljuk:

6. táblázat: A *directive* címkék felszólító alakokra vetített aránya

	nodir	sajáth	közv	meta	amb	szszerv	inter
álhír	0,3795	0,4891	0,0537	0,0288	0,0005	0,0288	0,0179
kontroll	0,6618	0,0742	0,1034	0,1107	0,0134	0,0207	0,0158
mind	0,4666	0,3611	0,0691	0,0541	0,0045	0,0263	0,0173

A 6. táblázatból egyértelműen látszik, hogy az álhírekben a felszólító alakokat sokkal többször használták saját hangú felszólításra, mint a kontroll szövegekben, ami alátámasztja a kiinduló hipotézisünket. A kontroll szövegekben a felszólító alakok ugyanakkor többször közvetítenek más által megfogalmazott felszólítást. A szöveg-

szervező és interakciós felszólítások aránya megegyezik, az egyéb nem felszólítást megvalósító alakok (*nodirectiva*) viszont a kontroll szövegekre jobban jellemzőek.

A *source* és a *target* jegyeket csak a saját hangú és a közvetített felszólítást kifejező felszólító alakok (dir) esetében jelöltük, ezeknek a címkéknek az előfordulási számait és a *saját hangú + közvetített* címkékhez viszonyított arányukat mutatja a 7. és 8. táblázat.

7. Táblázat: A *source* és a *target* címkével ellátott felszólító alakok száma a MedCollect korpuszban

	dir	**speaker**	**speaker+**	**vague**	**listener**	**listener+**	**inclusive**	**vague**
		source			**target**			
álhír	1000	839	160	1	498	44	457	1
kontroll	146	18	128	0	75	48	23	0
mind	1146	857	288	1	573	92	480	1

8. Táblázat: A *source* és a *target* címkéknek az olvasóra irányuló felszólításokra vetített aránya

	speaker	**speaker+**	**vague**	**listener**	**listener+**	**inclusive**	**vague**
	source			**target**			
álhír	0,8390	0,1600	0,0009	0,4980	0,0440	0,4570	0,0009
kontroll	0,1233	0,8767	0,0	0,5137	0,3288	0,1575	0,0
mind	0,7478	0,2513	0,0008	0,5000	0,0803	0,4188	0,0008

A *source* címkéknél a 8. táblázatban az a legszembetűnőbb jelenség, hogy az álhírekre sokkal jellemzőbb az, hogy a szöveg szerzője az egyedüli forrása a felszólításnak, a más forrású, vagy egy csoport általi felszólítások viszont a kontroll szövegekben szerepelnek nagyobb arányban. Ez összefügg azzal a 6. táblázatban látható megfigyeléssel, hogy az álhírekre jellemzőbb a saját hangú felszólítás, a kontroll szövegekre pedig a közvetített.

A direkt felszólítások között az álhírekben és a kontroll szövegekben is közel ugyanakkora, 50% az olvasót közvetlenül címző felszólítások aránya. Különbség mutatkozik viszont a *listener+* és az *inclusive* címkék között, az álhírekre jellemzőbbek az olvasót és a szöveg szerzőjét is célzó felszólítások (*viseljünk, figyeljünk*), a kontroll szövegekben viszont gyakoribbak az olvasóra mint egy csoport tagjára vonatkozó felszólítások (*mindenki viseljen*).

A felszólító alakokat tehát valóban különböző módon használják az álhírekben és a kontroll szövegekben. Az álhírekben gyakoribbak a felszólító alakok, a felszólító alakok között pedig gyakoribbak a saját hangú felszólítások. A tényleges felszólítások esetében a felszólítások címzettje is különbözhet a két csoportnál, az álhírek esetében

gyakoribbak a szerző és az olvasó közösségét kifejező inkluzív felszólítások, a kontroll szövegeknél pedig az olvasót egy közösség tagjaként érintő felszólítások.

A MedCollect egészségügyi álhírkorpusz felszólításannotált szavainak listája és az annotációk elérhetők a https://github.com/enyik-rg/directive_annotation github repozitóriumban.

6 Összegzés

Tanulmányunkban a MedCollect egészségügyi álhírkorpuszt és az azon elvégzett felszólításannotálást mutattuk be. Az álhírek terjedése elleni küzdelemben fontos szerepet játszik a hírolvasók médiatudatosságának növelése, amelynek az egyik módja az, hogy megmutassuk, milyen módszerekkel érik el az álhírek létrehozói a céljaikat.

A kiinduló hipotézisünk az, hogy az álhírek és az áltudományos szövegek rendelkeznek olyan nyelvi jegyekkel és nyelvhasználati stratégiákkal, amelyek alapján vagy amelyek kombinációi alapján egy szövegről gyanítható, hogy az álhír vagy áltudományos szöveg. Ezeknek a nyelvhasználati stratégiáknak a feltárása érdekében van szükség jó minőségű álhírkorpuszra, ami alapján azonosíthatjuk a nyelvhasználati stratégiákat és az azokat megvalósító nyelvi eszközöket.

Jelen tanulmányunkban a direkt felszólításokat vizsgáltuk. A MedCollect korpuszban kézi annotációval jelöltük a felszólító alakok különböző funkcióit, valamint az olvasóra irányuló tényleges felszólítások forrását és ezen felszólítások célpontját. Hipotézisünk a következő volt: az álhírek szignifikánsan több felszólító igealakkal megvalósított direktívát tartalmaznak mint a valódi hírek az olvasók felé irányuló erőteljesebb nyomásgyakorlás miatt. Hipotézisünket a korpusz adatai alátámasztották: az álhírekben a szöveg terjedelméhez viszonyítva is több felszólító alak jelent meg, mint a kontroll szövegekben, továbbá a felszólító alakok számához viszonyítva is sokkal nagyobb arányban szerepeltek a szöveg írójától származó közvetlen, eredeti felszólítások, míg a kontroll szövegekben a valódi felszólítást nem hordozó, vagy csak mástól származó, közvetített felszólítások voltak a jellemzőek.

Az olvasóra irányuló tényleges felszólítások tekintetében is találtunk különbséget az álhírek és a kontroll szövegek között: az álhírekre jellemzőbbek voltak az olvasót és a szöveg alkotóját egyaránt címzettnek tekintő inkluzív felszólítások. Ezzel szemben a kontroll szövegekben nagyobb arányban találtunk az olvasóra mint egy tulajdonság által meghatározott csoport tagjára irányuló felszólítást.

A felszólításannotálás eredményeit további kvalitatív és kvantitatív vizsgálat során kívánjuk felhasználni, illetve más nyelvhasználati stratégiák és nyelvi jegy annotálását is tervezzük a korpuszon.

Bibliográfia

Aijmer, K.: Corpus pragmatics: From form to function. In A. H. Jucker, K. P. Schneider, W. Bublitz (Szerk.), *Methods in Pragmatics* (o. 555–586). De Gruyter Mouton. (2018). https://doi.org/10.1515/9783110424928-022

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., Biemann, C.: A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. (2016). https://www.aclweb.org/anthology/W16-4011

Huszák D.: Példátlan információs háború zajlik Ukrajna körül – Elképesztő mennyiségű hazugság ömlik a világra. *Portfolio*. (2022, február 26). https://www.portfolio.hu/global/20220226/peldatlan-informacios-haboru-zajlik-ukrajna-korul-elkepeszto-mennyisegu-hazugsag-omlik-a-vilagra-529377

Indig B., Sass B., Simon E., Mittelholcz I., Kundráth P., Vadász N.: Emtsv – Egy formátum mind felett. In Berend G., Gosztolya G., Vincze V. (Szerk.), *MSZNY 2019, XV. Magyar Számítógépes Nyelvészeti Konferencia* (o. 235–247). Szegedi Tudományegyetem Informatikai Tanszékcsoport. (2019).

Islam, M. S., Sarkar, T., Khan, S. H., Mostofa Kamal, A.-H., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., Seale, H.: COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *The American Journal of Tropical Medicine and Hygiene*, *103*(4), 1621–1629. (2020). https://doi.org/10.4269/ajtmh.20-0812

Krekó P., Hercsel A., Szalai S.: *A koronavírus a kommunikációt is megfertőzi.* Index Kultrovat podcast 1. (2021.04.15). https://index.hu/kultur/2021/04/15/kreko-peter-tomegparanoia-podcast/

Németh T. E.: Álhírek, áltudományos nézetek nyelvészeti azonosítása. *Magyar Nyelv*, *119*(4), 490–496. (2023). https://doi.org/10.18349/MagyarNyelv.2023.4.490

Németh T. E., Nagy C. K., Németh Zs.: Ami a korpuszokból kimaradt: Rejtőzködő pragmatikai jelenségek. In Simon G., Tolcsvai Nagy G. (Szerk.), *Nyelvtan, diskurzus, megismerés* (o. 333–356). ELTE Eötvös Kiadó. (2020).

Rákosi, Cs.: *The Mask Denial Paradox. A new approach to the identification and analysis of pseudoscientific texts* [Kézirat]. Debreceni Egyetem. (2023a).

Rákosi Cs.: *A maszktagadó paradoxona*. Álhírek és áltudományos nézetek azonosítása a nyelvészet és a mesterséges intelligencia eszközeivel, Szeged. (2023b, november 17).

Searle, J. R.: Indirect speech acts. In P. Cole, J. L. Morgan (Szerk.), *Syntax and Semantics, Volume 3: Speech Acts* (o. 59–82). Academic Press. (1975).

Soltész K.: *Infodémia. Infojegyzet*. Országgyűlés Hivatala, Közgyűjteményi és Közművelődési Igazgatóság Képviselői Információs Szolgálat. (2023). https://www.parlament.hu/documents/10181/64399821/Infojegyzet_2023_21_infodemia.pdf

United Nations: *Our Common Agenda Policy Brief 8. Information Integrity on Digital Platforms*. United Nations. (2023). https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., Vincze, V.: E-magyar – A Digital Language Processing System. In N. Calzolari (Szerk.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (o. 1307–1312). European Language Resources Association (ELRA). (2018).