A CLARIN ParlaMint magyar korpusza

Üveges István[1,2], Ring Orsolya[2]

[1]Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
[2]Társadalomtudományi Kutatóközpont, Politikatudományi Intézet
uvegesistvan898@gmail.com
ring.orsolya@tk.hu

Kivonat Cikkünkben bemutatjuk CLARIN ParlaMint projekt keretében 2020 novembere és 2021 májusa között készült, a Covid19-járvány kommunikációjának vizsgálatára is alkalmas, egységes morfológiai és szintaktikai annotációt tartalmazó korpuszok között helyet kapó magyar nyelvű korpuszt, amely a magyar Parlamentben 2014 júniusa és 2020 decembere között elhangzott interpellációk és azonnali kérdések leiratait tartalmazza. Az eredeti leiratok a Magyar Országgyűlés honlapján[1] érhetőek el. Röviden ismertetjük a korpusz főbb szófaji statisztikáit, az alkalmazott (gépi) annotációs rétegeket, illetve bemutatunk néhány lehetséges alkorpuszokra való felbontást.
Kulcsszavak: parlamenti korpusz, clarin, msd, xml, Covid19

1. Bevezetés

A parlamentek a politikai kommunikáció fontos helyszínei, ahol a választott képviselők megvitatják a benyújtott törvényjavaslatokat és más országos jelentőséggel bíró ügyeket. Az itt elhangzó beszédek általában előre megtervezett beszédaktusok, mivel a képviselők kiemelt célja, hogy meggyőzzék a hallgatóságot és megszerezzék támogatásukat. A parlamenti viták jegyzőkönyveinek egyedi tartalma, szerkezete és nyelvezete fontos forrásai a társadalomtudományi- és nyelvészeti kutatásoknak. A politikai kommunikáció korpuszokon és NLP módszereken alapuló kutatása az elmúlt időszakban kiemelt jelentőséget kapott, de megjelent tanulmányok legtöbbször a politikusok médiában és közösségi médiában megjelenő megnyilatkozásait elemzik (Gollust és mtsai, 2020; Mariani és mtsai, 2020; Aparicio és mtsai, 2021; Wang és mtsai, 2021; Rufai és Bunce, 2020).

A parlamenti viták leiratai lényegében a beszélt nyelv ellenőrzött és szabályozott körülmények között készült átiratai, melyek szabadon elérhetőek, mivel az információszabadságról szóló törvény alapján nem vonatkoznak rájuk a szerzői jogi vagy a személyes adatok védelmére vonatkozó jogszabályok. Éppen ezért az utóbbi években több nemzetközi projekt keretében készült és készül korpusz parlamenti felszólalásokból[2].

[1] https://www.parlament.hu/

[2] Ilyen például a CLARIN https://www.clarin.eu/, a Comparative Agendas https://www.comparativeagendas.net/ vagy az OPTED https://opted.eu/ projekt

A CLARIN kutatási infrastruktúra keretében lezajlott ParlaMint projekt[3] célja egységesen kódolt, ezáltal jól összevethető többnyelvű, nyelvészeti annotációval ellátott korpuszok létrehozása volt. A projekt keretében 17 ország parlamenti felszólalásai kerültek feldolgozásra, összesen mintegy 500 millió token terjedelemben, amelyből a magyar korpusz mintegy 1,019,576 token.

A Magyar Országgyűlésben elhangzott interpellációkból és azonnali kérdésekből politikatudományi felhasználásra már készült korpusz a Hungarian Comparative Agendas Project keretében[4] amely ugyan nem tartalmaz nyelvészeti és szintaktikai annotációt, azonban minden tekintetben alkalmas volt arra, hogy a szükséges nyelvészeti és szintaktikai elemzésekkel és minimális metaadat kiegészítésekkel a nemzetközi korpusz részévé váljon, ezzel kapcsolódási lehetőséget teremtve a politikatudományi és nyelvtudományi célra épült korpuszok között.

A ParlaMint projekt során létrejött korpuszok időbeli eloszlása és nagysága is különböző. Néhány alapelvtől eltekintve a résztvevő kutatócsoportok döntésén alapult, hogy mely parlamenti beszédtípusokat, milyen időintervallumban dolgoznak fel. A létrehozott korpuszokban a 2019 novembere után keletkezett szövegek a Covid19-korpuszba, míg a korábbi szövegek a referenciakorpuszba kerültek. A referenciakorpuszok időhatára alkalmazkodhatott az egyes országok parlamenti ciklusaihoz, de a kezdődátuma nem lehetett 2015 utáni. A korpuszok CLARIN TEI XML séma [5] szerint készültek, emellett egységes szemléletű nyelvészeti és szintaktikai feldolgozáson esetek át.

Mivel a parlamenti beszédek leiratainak egyik fontos jellemzője, hogy közvetlenül reagálnak a bekövetkező eseményekre, így például a jelenlegi Covid19-világjárványra, a 17 nyelven létrehozott korpuszok az adatok szinkron és diakronikus összehasonlításán keresztül alkalmasak a járványhoz kötődő kommunikáció többnyelvű kontextusban történő vizsgálatára.

A korpuszok kiterjedt metaadat-struktúrával rendelkeznek a felszólalókról (név, nem, pártállás, képviselői státusz) és a parlamenti ülésekről, emellett minden beszéd mellett megtalálható előadójának aktuális szerepe (elnök, rendes előadó) is. A beszédek emellett tartalmaznak az elhangzott szövegekre vonatkozó olyan megjegyzéseket is mint például a közbeszólások, bekiabálások vagy a taps. A korpuszok letölthetőek a CLARIN.SI repozitóriumból[6] és elérhetőek `noSketchEngine`-en keresztül[7]. A repozitóriumban elérhetők a korpuszvalidáláshoz használt `XLST` és `Perl` állományok, amelyek hasznosak lehetnek a TEI XML fájlok tovább alakítása esetén.

A tanulmány a következők szerint épül fel; a 2. fejezet a vizsgált szövegtípusokat ismerteti röviden, majd a 3. fejezetben a magyar korpusz főbb jellemzőit mutatjuk be, míg a 4. fejezet az XML sémában elhelyezett nyelvészeti annotációt, és az ennek elkészítéséhez használt eszközöket tárgyalja. A 5. fejezet-

[3] https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora
[4] https://cap.tk.hu/hu
[5] https://github.com/clarinsi/TEI-schema
[6] https://www.clarin.si/info/about-repository/
[7] https://www.clarin.si/noske/parlamint.cgi/

ben részletesen kitérünk a szövegek nyelvészeti annotációinak néhány fontosabb TEI XML specifikus jellemzőjének ismertetésére. A tanulmányt ezt követően rövid konklúzió zárja.

2. Az interpelláció és azonnali kérdés, mint a parlamenti ellenőrzés eszközei

Az interpelláció és az azonnali kérdés a képviselők által gyakorolható hagyományos parlamenti ellenőrzési eszköz. Különbség közöttük a címzettek körében és a tárgyalási rendjükben van.

Az Alaptörvény 7 cikk (2) bekezdése szerint az országgyűlési képviselők joga, hogy interpellációt intézzenek a Kormányhoz és a Kormány tagjához a feladatkörükbe tartozó bármely ügyben [8]. Az interpelláció során a képviselő szóban ismerteti az interpelláció szövegét, majd a válasz és a képviselői viszontválasz következik. Végül a plenáris ülés szavaz arról, hogy elfogadja-e a választ, avagy elutasítva azt a kérdésről egy bizottsággal jelentést készíttet. Az interpelláció címzettje csak a kormány vagy annak valamely tagja lehet. Az interpelláció szövegét napokkal elhangzása előtt be kell nyújtani (Magyar, 2018).

Az 1994-ben bevezetett azonnali kérdéseket a frakcióvezetők terjesztik be, majd képviselők mondják el. Minden héten legalább hatvan perc áll rendelkezésre az azonnali kérdésekre, és minden képviselőcsoportot megilleti a jog ezalatt legalább egy azonnali kérdés ismertetésére, melyeket legalább az ülés megkezdése előtt hatvan perccel be kell nyújtani.

3. A magyar korpusz jellemzői

Az elkészült magyar korpusz egy lezárt (2014-2018) és a jelenleg is folyamatban lévő (2018-) parlamenti ciklusban elhangzott valamennyi interpelláció és azonnali kérdés szövegét tartalmazza. A szövegeket web scraping segítségével kerültek letöltésre a Magyar Országgyűlés honlapjáról, az alapvető metaadatokkal együtt, a Hungarian Comparative Agendas Project keretében.

Az interpellációkból és az azonnali kérdésekből így létrehozott adatbázisban szereplő legfontosabb változók az alábbiakra terjednek ki: az interpelláció címe, az interpelláció beterjesztőjének neve, az interpelláció beterjesztésének időpontja, az interpellációk közpolitikai tartalma, a válaszadó neve és az Országgyűlés döntése a miniszteri válasz elfogadásáról. Az azonnali kérdések esetében pedig azok címe, a beterjesztő neve, az azonnali kérdés közpolitikai tartalma, az azonnali kérdés címzettjének neve, valamint a beterjesztés időpontja. A CLARIN ParlaMint projekt során ezen adatbázisokat és a hozzá tartozó szövegállományt alakítottuk CLARIN TEI XML formátumú korpusszá.

[8] https://njt.hu/jogszabaly/2011-4301-02-00.11

3.1. Kereshetőség

Az online keresőfelület[9] lehetővé teszi többek között:

- konkordancia készítését lemma, frázis, szóalak, karakter vagy CQL alapon,
- szűrés kontextusra (+/- 15 token távolságig),
- a korpusz többféle felosztását, például parlamenti ciklusok vagy a koronavírus járvány kitörését megelőző, és az azt követő időszak felszólalásaira,
- keresést adott frakció hozzászólásaiban illetve nemek szerint is (a részletesebb beállítási lehetőségeket az 1. ábra szemlélteti).

1. ábra. A korpusz keresőfelületének néhány beállítási lehetősége.

[9] https://www.clarin.si/noske/parlamint.cgi/first_form?corpname=parlamint21_hu;align=

Ahogyan korábban kifejtettük a ParlaMint projekt célkitűzése szerint a létrejövő korpusz fő fókuszában az állt, hogy a Covid19 járvány megjelenését és hatásait a nemzeti parlamentekben elhangzó felszólalásokban egyszerűen követhetővé és vizsgálhatóvá tegye, ennek kapcsán tehát a felület natívan kezeli a korpusz ilyetén felosztását. Hasonlóan egyszerűen elvégezhető például az ellenzék - kormánypártok felosztás (a megfelelő pártok együttes kijelölésével), vagy akár a fentiek kombinálása a megfelelő parlamenti ciklus kijelölésével. Mindezeknek köszönhetően a szövegek alkalmasak lehetnek például az egyes pártoknak a koronavírussal összefüggő kommunikációja vizsgálatára, vagy akár konkrét képviselők felszólalásainak összevetésére is.

3.2. Leíró statisztikák

Ahogyan már említettük, a magyar CLARIN ParlaMint korpuszt hozzávetőlegesen 1 millió tokenes szövegállomány alkotja. Az 1. táblázat néhány lehetséges felbontás szerint mutatja be a képezhető részkorpuszok szófaji statisztikáit.

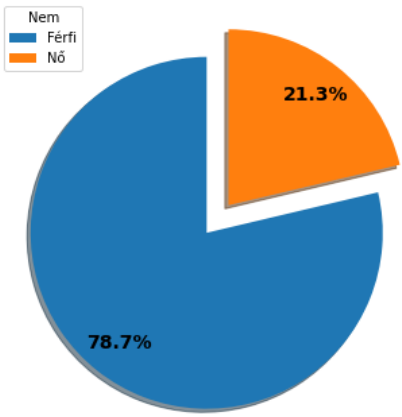
	post-covid	**pre-covid**	**1.ciklus**	**2.ciklus**	**ellenzék**	**kormány**
token	198.930	820.646	635.791	383.785	798.449	221.127
mondat	11.030	43.968	33.995	20.983	44.189	13.654
NOUN	43.923	187.066	145.129	85.850	178.393	57.349
ADJ	21.585	92.898	71.504	42.962	81.859	30.612
PRON	14.492	56.721	44.013	27.218	59.122	10.608
CONJ	8.895	36.664	28.223	17.325	35.034	9.739
NUM	4.123	21.262	17.254	8.143	19.427	5.543
VERB	22.354	86.680	66.888	42.149	87.269	20.257
ADV	15.635	62.682	48.497	29.899	64.169	12.696
PROPN	5.958	24.943	18.964	11.930	23.570	8.221
ADP	2599	11.645	9.031	5.221	10.524	3.401
AUX	1	1	1	1	1	1
DET	20.420	84.150	65.221	39.324	82.164	24.341
INTJ	273	873	670	475	969	170
PART	761	2.769	2.110	1.423	2.858	643
PUNCT	31.734	128.756	100.218	60.213	128.624	32.892
SCONJ	6.158	23.454	17.999	11.620	24.384	4.627
SYM	0	5	5	0	4	1
X	18	76	62	31	72	25

1. táblázat. Különböző szófaji címkék számossága a CLARIN ParlaMint korpusz néhány felbontása esetén.

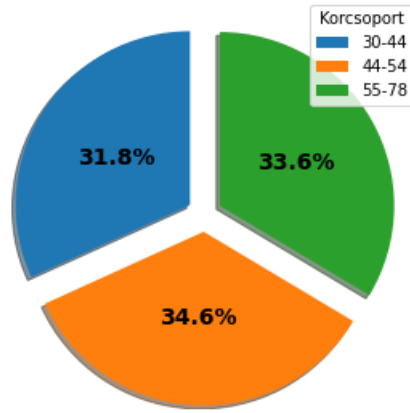
Tekintettel arra, hogy a korpuszba kerülő szövegek időarányosan kerültek kiválogatásra, így a Covid19 kitöréséhez képest kialakítható részkorpuszok aszimmetrikusan alakulnak a járványt megelőző időszak javára. Az ellenzék - kormány felosztás hasonló mértékű aszimmetriája (a részkorpuszok tokenszámát tekintve) az interpellációk / azonnali kérdések természetének tudható be; itt ellenzéki

pártok intéznek kérdést a kormányhoz, majd a válasz után szintén az ő viszontválaszuk következik.

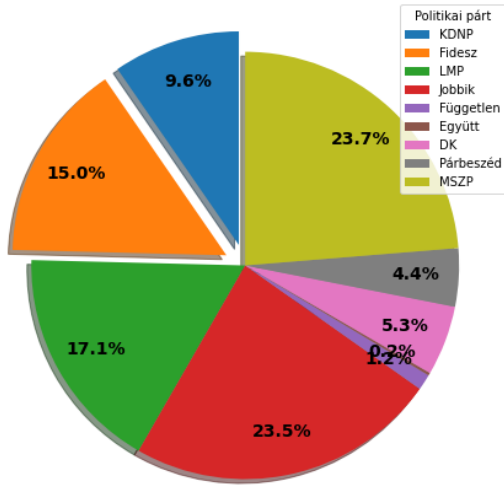
A 2. ábra az XML-ben kódolt metaadatok alapján kiválogatott felszólalások számarányát mutatja három lehetséges bontásban.



(a) Nemek szerint



(b) Korcsoportok szerint



(c) Politikai pártok szerint

2. ábra. Felszólalások arányai a CLARIN ParlaMint magyar korpuszában.

Nemek szerinti csoportosítva a felszólalásokat azt látjuk, hogy jelentős túlsúly mutatkozik a férfi képviselők javára a nőkkel szemben (78,7% a 21.3%-kal szemben). Érdemes megemlíteni, hogy a női képviselők aránya a parlamentben az első

(2014-2018-ig tartó) parlamenti ciklusban, amelyet a korpusz tartalmaz 10.1% körül alakult, míg a második ciklus (2018-) esetében ez az arány 12,6% körül alakul[10].

Habár ez a 3. legalacsonyabb arány Európában (az Európai Unió átlaga nagyjából 30% körül mozog), a női képviselők hozzászólásainak a nők számarányához mért közel kétszeres aránya arra enged következtetni, hogy igen aktívan részt vesznek a parlamenti üléseken zajló politikai diskurzusban.

Korcsoportok szerint osztályozva a képviselőket a két ciklus átlagában jól kirajzolódik egy öregedő korfa; a jelen ciklusban mindösszesen 2 fő 30 év alatti képviselő rendelkezik mandátummal, és ez a szám a megelőző ciklusban is mindössze 4 fő volt. A leginkább jellemző a felszólalók között az 50 év körüli életkor volt.

Politikai pártok tekintetében a legaktívabbnak egyértelműen az MSZP képviselői tekinthetők; az 1. ciklusban a képviselői mandátumok 14,5%-a, a 2. ciklusban pedig 7,5% -a volt a párt birtokában, ezzel szemben ők adták az összes felszólalás mintegy 23,5% -át. A legkevésbé aktív ezzel szemben egyértelműen a Fidesz volt, akik 58,2% illetve 58,8%-nyi mandátumukhoz a hozzászólások 15%-ával rendelkeznek a két ciklus összesítésében. Kormánypárt - ellenzék szerint polarizálva a számosságokat 24,6% adódik a 75,4% ellenében, amely a mandátumok eloszlásának (66,83% a 33.17% ellenében) közel fordítottja. Ezek az arányok mind az MSZP (ellenzék) - Fidesz (kormány), mind az ellenzék - kormány viszonylatban egyértelműen a beszédtípusok bevezetőben említett jellegzetességével magyarázhatóak.

3.3. Lexikai alapú doménhasonlóság

Annak érdekében, hogy képet kaphassunk a felépített korpusz hasonlóságáról más domének szövegeihez viszonyítva, a korpusz szövegét a Jaccard-távolság metrika felhasználásával összevetettük a Szeged Treebank (Vincze és mtsai, 2010) 6 részkorpuszával, amelyek tartalma;

- iskolai fogalmazások,
- szépirodalmi szövegek,
- számítástechnikai szövegek,
- újsághírek,
- jogi szövegek,
- valamit üzleti rövidhírek

közül került ki. A Jaccard-távolság alapját a vizsgált szövegek szókészlete adja; arról ad visszajelzést, hogy az összevetett szövegek esetében mekkora arányú a közös szókincs, értéke 0 és 1 között változik, ahol 1 a tökéletes egyezést jelenti, 0 pedig azt, hogy a két mért szöveg szókincse diszjunkt halmazt alkot[11]. A mért távolságokat a 2. táblázat mutatja be.

[10] Forrás: Eurostat (https://ec.europa.eu/eurostat/databrowser/view/sdg_05_50/default/-table?lang=en)

[11] $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, ahol A és B a két szöveg szavaiból (pl. lemmák) képzett halmazok.

	Jaccard távolság
Üzleti rövidhírek	0,8390
Számítástechnika	0,8386
Szépirodalom	0,8450
Újsághírek	**0,7551**
Iskolai fogalmazások	0,8354
Jogi szövegek	**0,8701**

2. táblázat. A CLARIN ParlaMint korpusz Jaccard-távolsága a Szeged Treebank egyes részkorpuszaitól.

Ez alapján az elkészült korpusz legtávolabb az újsághírek szókincsétől helyezkedik el, míg a legnagyobb átfedést a Szeged Korpuszban a jogi doménbe sorolt szövegekkel mutatja. A távolságok megoszlásában az újsághírek 0.75-ös értéke egyértelműen szélsőségesnek számít, de kiugrónak tekinthető a legközelebbi részkorpusz (jogi szövegek) 0.87-es értéke is, tekintettel arra, hogy a fennmaradó 4 részkorpusz távolsága meglehetősen homogén (rendre 0.83 - 0.84 körül ingadozik).

A jogi szövegekkel vett legnagyobb hasonlóság várható volt, tekintettel arra, hogy a parlamenti felszólalások témája sok esetben a jogalkotási folyamathoz kötődik, ami eszerint tehát világosan leképeződik a használt szókincseben is. Az újsághírektől vett (a többi részkorpuszhoz képest) kiugróan nagy távolság feltehetőleg azok témaválasztásbeli változatosságával magyarázható; a kevéssé egységes topikok széttartóbb szókincset eredményezhetnek.

4. Annotációs rétegek

Ahogyan említettük, annak érdekében, hogy a korpuszban helyet kapó valamennyi nyelvű leiratok összevethetőek maradjanak, azokat egységes nyelvészeti annotációval kellett ellátni. A korpuszból a munka során két változat készült, amelyek közül a nyelvészetileg elemzett korpuszvariánsnak az alábbi annotációkkal kellett rendelkeznie:

- Univerzális Dependencia (UD) szerinti szintaktikai elemzés (Zeman és mtsai, 2020)
- az egyes tokenekhez a megfelelő MSD kód hozzárendelése (Erjavec, 2012)
- a mondatokban szereplő névelemek taggelése.

Tekintettel arra, hogy magyar nyelvre egyben egyetlen elemző sem biztosítja mind a három fenti standard szerinti kimenetet, ezért az előelemzés három különböző eszköz kimenetének egyesítésével volt csak megoldható. Az Univerzális Dependencia szerinti függőségi nyelvtani címkézést a UDPipe 2.0 elemző (Straka, 2018) REST API -ként elérhető szolgáltatásával valósítottuk meg, az MSD kódolást a magyarlanc régebbi, 2.0-ás változata (Zsibrita és mtsai, 2013), a

névelemek azonosítását és csoportokra bontását pedig a Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportjában fejlesztett névelem-felismerő (Szarvas és mtsai, 2006) biztosította.

A fenti eszközök mindegyike az előfeldolgozás lépéseként tokenizálja és mondatokra szegmentálja a kapott szöveget, azonban ezek a felbontások az egyes elemzők esetében nem feltétlenül esnek egybe. Ennek következtében az egyes kimenetek egyesítése során szükséges volt kiválasztani egy "etalont", amelybe a többi címkekészlet elemeit integráljuk. Erre a célra (lévén mind közül ez a legkorszerűbb) a UDPipe elemző kimenetét választottuk, más szóval az ezáltal előállított tokenekhez kerestünk a másik két elemző kimenetében megfelelő címkézést. A címkéket akkor tekintettük megfeleltethetőnek, és egyesítettük egy közös formátumba, amikor a tokenizálás azonos eredményt hozott valamennyi eszköz esetében.

A kimenetek összeillesztése automatikusan történt. Azokban az esetekben, amikor a magyarlanc 2.0 vagy a névelem-felismerő címkézése az eltérő tokenizálás miatt nem volt konzisztens a UDPipe kimenetével, az ilyen módon hiányzó címkék helyére technikai adatokat helyeztünk el, indikálva, hogy a megfelelő kimenetek nem voltak egyesíthetők. Az eredeti teljes szövegmennyiség egy kisebb részhalmazán végzett kézi ellenőrzés alapján ilyen hibák az összes elemzett szövegnek mintegy néhány százalékát érintették.

5. TEI XML integráció

A fenti lépések során kinyert morfológiai és szintaktikai információkat a munka következő fázisában a projekt alapját képező Parlamint TEI XML sémába[12] illesztettük. A séma teljes leírása messze meghaladná a jelen tanulmány kereteit, így itt csak néhány fontosabb elem ismertetésére térünk ki.

Amennyiben minden fázis sikeresen végbemenet, a morfológiai annotációt egy `<w>` -tag zárta egységbe, amelynek attribútumai és értéke a következők szerint alakult:

```
(1)  <w pos="Pd3-sn"
        lemma="olyan"
        msd="UPosTag=ADJ|Case=Nom|Degree=Pos|Number=Sing"
        xml:id="IC7_157_2.2.2.1">
        Olyan
     </w>
```

Fontos kiemelni, hogy a használt XML séma elvárásainak megfelelően az MSD kód a `pos` attribútum értékeként, a UDPipe által meghatározott morfológiai jegyhalmaz pedig az `msd` attribútum értékeként jelent meg. Az XML tag-ben ezen felül még az adott tokenhez a UDPipe által rendelt `lemma` szerepelt a neki megfelelő attribútum értékeként, az `xml:id` pedig a tokennek az

[12] https://clarin-eric.github.io/parla-clarin/

adott nap parlamenti felszólalásai között elfoglalt helyét jelölte ki (a fájl azonosítója, pl.: IC_157_2, majd pontokkal elválasztva a fájlon belüli felszólalás sorszáma, azon belül a mondat és a mondaton belül a konkrét token sorszáma). A névelemek jelzésére a `<name>` tag szolgált, amely több tokenes névelemek esetén magában foglalta valamennyi tokent (`<w>`), és amelynek `type` attribútuma a névelem típusát jelölte (ORG - organization, PER – Person, LOC - Location vagy MISC - Miscellaneous):

```
(2) <name type="ORG">
        <w lemma="kuria"
            msd="UPosTag=PROPN|Case=Sub|Number=Sing"
            xml:id="IC7_165_2.1.4.3">
            Kuriara
        </w>
    </name>
```

Az egyes mondatok tokenenként kódolt morfológiai információi után a szintaktikai elemzés `linkGrp` tagen belül foglalt helyet; minden token egy önálló `link` taget kapott, amelynek `ana` attribútuma adja meg az UD szintaktikai élcímkét, a `target` attribútum pedig az él kiindulását és érkezését:

```
(3) <link ana="ud-syn:det"
    target="#IC7_164_2.1.3.5 #IC7_164_2.1.3.1"/>
```

A fenti példában a `#IC7_164_2.1.3.1` jelenti az aktuális token számát; ez lesz az él kiindulása, a `#IC7_164_2.1.3.5` pedig az él érkezési tokenjét kódolja, így ezek végig követesével a teljes szintaktikai fa visszafejthető. Az UD elemzésben a mondat fejének tekintett `root` komponens annyiban speciális, hogy őseként a mondat azonosítója van megjelölve, tokenszám nélkül (a fenti példa esetében: `#IC7_164_2.1.3`).

A korpuszban emellett jelölve lettek a parlamenti leiratozók által feltüntetett különféle hanghatások (mint például taps, csengetés). Ezeket a `kinesic` tagek hivatottak kódolni az egyes megszólalások teljes szövege után (vagyis jelzésük nem a felszólalás alatti elhangzás valós ideje szerint történt), pl.:

```
(4) <kinesic type="vocal">
        <desc>(Zaj. – Az elnok csenget.)</desc>
    </kinesic>
```

A korábban már említett központi `Git` repozitóriumban a korpusznak két xml variánsa található meg. Ezek között a fő különbség, hogy míg az egyik a nyelvészetileg annotált, a fentieknek megfelelő tag-eket magukban foglaló fájlokat tartalmazza (.ana.xml kiterjesztéssel), addig a másik változatban a hozzászólások hagyományos szöveges formában, néhány hozzájuk rendelt metaadattal szerepelnek, felszólalások szerint bontva.

Ezek esetében a metaadatok közül direkten a felszólaló neve érhető el (pl.: `<note>DR. TÓTH BERTALAN (MSZP):</note>`), a többi metaadat a felszólalás azonosítójához rendelve érhető el a ParlaMint-HU.xml fájlba szervezve. Egy-egy képviselőhöz például az 1. fejezetben említett adatok a következők szerint kereshetők:

```
(5) <person xml:id="TiborBana">
        <persName>
            <forename>Tibor</forename>
            <surname>Bana</surname>
        </persName>
    <sex value="M">Ferfi</sex>
    <birth when="1985">1985</birth>
    <affiliation role="member" ref="#party.FUGGETLEN" />
    <affiliation role="MP" />
    </person>
```

A `<teiHeader>` mindkét esetben tartalmazza például az egyes fájlok további metaadatait (mint amilyen az alkalmazott tag-ek száma), így azok különösen hasznosak lehetnek leíró statisztikák készítéséhez. Külön fájlba szervezve, szintén a `<teiHeader>` tag tartalmazza a teljes korpuszra és metaadatokra vonatkozó összesített információkat is.

6. Összegzés

Cikkünkben röviden bemutattuk a CLARIN kutatási infrastruktúra ParlaMint projekt keretében készült, magyar nyelvű parlamenti felszólalásokat tartalmazó korpuszát. A korpuszban foglalt szöveganyag, illetve az elkészült morfológiai és szintaktikai annotáció, illetve a korpusz metaadatai lehetőséget teremtenek például a különböző parlamenti frakciók kommunikációjának elemzésére, a képviselők megnyilatkozásainak vizsgálatára a Covid19-et megelőző és az azt követő időszakban, vagy éppen a képviselők felszólalásainak kor-, nem- és pártállás szerinti bontásban történő elemzésére is.

A jövőben a korpusz kiegészítését tervezzük más parlamenti beszédtípusokkal, valamint további parlamenti ciklusok felszólalásainak szöveganyagával, ezáltal teret biztosítva széles körű nyelvészeti és társadalomtudományi vizsgálatoknak.

Köszönetnyilvánítás

A publikációban szereplő kutatást, amelyet a Társadalomtudományi Kutatóközpont valósított meg, az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia

Nemzeti Laboratórium keretében. Külön köszönet illeti a Társadalomtudományi Kutatóközpont Comparative Agendas Project kutatócsoportjának tagjait és gyakornokait a felhasznált korpuszok előkészítéséért.

The research was supported by the European Union's Horizon 2020 research & innovation programme under Grant Agreement no. 951832.

The research was supported by CLARIN ERIC ParlaMint Project.

Hivatkozások

Aparicio, J.T., de Sequeira, J.S., Costa, C.J.: Emotion analysis of portuguese political parties communication over the covid-19 pandemic. In: 2021 16th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–6. IEEE (2021)

Erjavec, T.: Multext-east: morphosyntactic resources for central and eastern european languages. Language Resources and Evaluation 46(1), 131–142 (2012), http://www.jstor.org/stable/41486069

Gollust, S.E., Nagler, R.H., Fowler, E.F.: The emergence of covid-19 in the us: a public health and political communication crisis. Journal of health politics, policy and law 45(6), 967–981 (2020)

Magyar, Z.: A parlamenti ellenőrzés eszközei az országgyűlés gyakorlatában. Parlamenti Szemle 2, 125–150 (2018)

Mariani, L.A., Gagete-Miranda, J., Retti, P.: Words can hurt: How political communication can change the pace of an epidemic. Covid Economics 12, 104–137 (2020)

Rufai, S.R., Bunce, C.: World leaders' usage of twitter in response to the covid-19 pandemic: a content analysis. Journal of public health 42(3), 510–516 (2020)

Straka, M.: UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 197–207. Association for Computational Linguistics, Brussels, Belgium (Oct 2018), https://aclanthology.org/K18-2020

Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In: Todorovski, L., Lavrac, N., Jantke, K.P. (szerk.) Discovery Science. pp. 267–278. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010), http://www.lrec-conf.org/proceedings/lrec2010/pdf/465$_P$aper.pdf

Wang, Y., Croucher, S.M., Pearson, E.: National leaders' usage of twitter in response to covid-19: A sentiment analysis. Frontiers in Communication p. 183 (2021)

Zeman, D., Nivre, J., Abrams, M.: Universal dependencies 2.6 (2020), http://hdl.handle.net/11234/1-3226, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A tool for morphological and dependency parsing of hungarian. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. pp. 763–771. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (9 2013), https://aclanthology.org/R13-1099