Eötvös Loránd Tudományegyetem

Társadalomtudományi Kar

ALAPKÉPZÉS

Automatizált szövegelemzés no-code eszközökkel: A Meaning Extraction Helper és az AntConc bemutatása a 2022-es hazai online kivándorlás-diskurzus vizsgálatával

Konzulens:
dr. Németh Renáta

Készítette:
Kovács Anna Krisztina
ZLGHLC
szociológia szak

2024. április

Tartalomjegyzék

Absztrakt

Az elmúlt évek során a témában készült tanulmányok száma alátámasztja, hogy az automatizált szövegelemzésnek egyre nagyobb szerepe van az empirikus társadalomkutatáson belül. Ebben a szakdolgozatban két olyan automatizált szövegelemzési eszköz kerül bemutatása, amelyek nem igényelnek programozói tudást, de releváns szociológiai kutatási kérdéseket képesek megválaszolni. Az eszközök lehetőségei és limitációi korábbi tanulmányok ismertetésével, valamint a dolgozatban végzett példa kutatáson keresztül kerülnek szemléltetésre. Az elemzésben a 2022-es országgyűlési választásokat követő, laikus közvéleményben megjelenő kivándorlás-diskurzusokat tekintem át. A Meaning Extraction Method alkalmazásával a diskurzusok fő témái, az Antconc használatával pedig a leggyakrabban használt szavak kontextusa kerül bemutatásra.

1. Bevezetés

Az információs társadalom korszakában egyre növekszik az elérhető adatok mennyisége és diverzitása. A társadalomtudományok területén ez új kapukat nyit meg, amely olyan attitűdök feltárását teszi lehetővé, amire korábban nem volt példa. Ebben a változó környezetben a hagyományos kutatási módszerek mellett egyre nagyobb igény van a technológiai innovációkra és az automatizációra. Az utóbbi években, hogy e hatalmas mennyiségű szöveg feldolgozásra kerülhessen, a szövegelemzés és az adatelemzés területén rendkívül gyors fejlődés volt megfigyelhető. Ahhoz azonban, hogy e lehetőségek kihasználására kerüljenek, olyan tudás szükséges, ami jelenleg nem része a társadalomtudósok eszköztárának. Ságvári (2017) véleménye szerint „egyre nagyobb igény lesz olyan alapvetően társadalomtudományos indíttatású szakemberekre, akik az elméleti felkészültségük és a nyilvánvalóan elvárható társadalmi érzékenységük mellett az új típusú adatok által megkívánt módszertani ismeretekkel és gyakorlati (programozói) tudással is rendelkeznek.” A nagy mennyiségű szöveges adat feldolgozása és elemzése tehát komplex, több tudományterületen átívelő tudást igényel.

Mindezzel párhuzamosan azonban olyan kezdeményezések is megjelentek, melyek az automatizált szövegelemzés hozzáférhetőségének növelésére irányulnak. Olyan eszközöket fejlesztettek, melyek könnyen elsajátíthatóak, és lehetővé teszik több ezer szöveg gyors

feldolgozását másodpercek alatt, anélkül, hogy a felhasználónak mélyreható programozói ismeretekkel kellene rendelkeznie. Az ilyen eszközök tehát bárki számára hozzáférhetővé teszik az automatizált szövegelemzést, hiszen a programozás, mint belépési küszöb, már nincs jelen. Maga a módszer így szélesebb körben alkalmazható kutatásokhoz, de segíthet az oktatásban is. Az eszközök használata ugyanis egyszerű, könnyen tanulható, mégis alkalmas nagyobb korpuszok hatékony elemzéséhez. E dolgozatban két olyan automatizált szövegelemző eszköz - a Meaning Extraction Helper és az Antconc - kerül bemutatásra és tesztelésre, melyek használata egyszerű, kódolást nem igényel, mégis alkalmazhatóak társadalomtudományos témájú kutatások módszertani eszközeként. A dolgozat továbbá hangsúlyt fektet az eszközök, valamint maga a no-code megoldás hiányosságaira is. Az eszközök limitációi ugyanis befolyásolhatják a szövegek elemzésének eredményeit is. Arra keresem a választ, hogy milyen mértékben képesek a vizsgált no-code eszközök megfelelő alternatívát nyújtani a programozás alapú szövegelemzési megoldásokhoz képest. Az említett eszközök működése a kivándorlás, valamint az azzal kapcsolatban megjelenő diskurzusok témáján keresztül kerül bemutatásra és tesztelésre.

A Magyarországról történő kivándorlás témája kevéssé kutatott, ugyanakkor befolyásolja az ország gazdasági és társadalmi életét, demográfiai változásokat hoz és szorosan kapcsolódik politikai kérdésekhez is. Az online sajtóban ezzel kapcsolatban közölt cikkek segíthetnek a kivándorlással kapcsolatos témák mélyebb megértésében. Emellett, a közösségi médiában erről folytatott diskurzusok segíthetnek megérteni az emberek attitűdjeit és érzéseit e témával kapcsolatban. Az ilyen felületek ugyanis fontos csatornái a laikus közvélemény megjelenítésének. A közösségi média gyakran szolgál platformként az embereknek, hogy megosszák gondolataikat és érveljenek döntéseik mellett, vagy éppen kifejezzék aggodalmaikat, elégedetlenségüket, haragjukat.

Ez utóbbiak fokozottan érvényesek a 2022-es országgyűlési választások idejére. A választásokkal járó általános érzelmi túlfűtöttség mellett a társadalmi és politikai feszültségek is kiéleződtek ebben az időszakban (Bíró-Nagy, 2022). Ekkor tehát a kivándorlás-diskurzusok elemzése is kiemelten fontos. A választásokhoz kapcsolódó érzelmi túlfűtöttség és a politikai feszültségek kiéleződése ugyanis újabb kontextust teremtett e diskurzusok számára. Szakdolgozatom második kutatási kérdése, hogy milyen fő témák és diskurzusok jelentek meg a Magyarországról történő kivándorlással kapcsolatban a laikus közvéleményben és az online

sajtóban a 2022-es év második felében. Ezt a kérdést a két automatizált szövegelemzési eszköz segítségével vizsgálom, tesztelve azok funkcióit.

A dolgozat első részében ismertetésre kerülnek az automatizált szövegelemzés szükségességének okai, valamint a kódolást csak részben, vagy egyáltalán nem igénylő megoldások jelentősége. Emellett bemutatásra kerül a kutatási kérdések relevanciája és háttere. A dolgozat második részében áttekintem a felhasznált eszközök korábbi tudományos alkalmazását, és részletezem azok funkcióit. Az ezt követő módszertani ismertetés után bemutatom az elemzés eredményét, amely mind a vizsgált kivándorlás-diskurzusokat és az abban megjelenő témákat, mind az eszközök használatával kapcsolatos megfigyeléseket magában foglalja. Végül, az eszközök használatával vizsgálható, más releváns kutatási irányokat vetek fel.

2. Elméleti áttekintés

2.1. Az információs társadalom

A gyors technológiai fejlődés, az internet megjelenése majd általános használata egyre több információt termel. Az információ létrehozása, terjesztése és manipulációja pedig nem csak az egyén életére, munkájára és az egyének közti kommunikációra van hatással, hanem befolyásolja a gazdaságot, a kultúrát és a társadalmat is (Mayer-Schönberger és Cukier, 2013). Az információs társadalomban tehát az információ a legfontosabb társadalomszervező tényező. Mindemellett azonban a társadalomkutatás empirikus adat-elérésére nézve is jelentős hatással van. Ugyanis, az „internetre került társadalmi folyamatok minden történése nyomot hagy maga után” (Csepeli, 2015, p.173).

A weboldalak, online könyvtárak és online híroldalak napi szinten közölnek új információkat. A web 2.0 (O’Reilly, 2009) azt a változást jelzi, amikor a felhasználók már nem csak az internetes tartalmak fogyasztói, hanem azok gyártói is lettek. Ekkor jelennek meg a blogok, a képmegosztó oldalak és a nyílt közösség által fejlesztett oldalak mellett az olyan közösségi média platformok is, mint a Facebook (Han, 2011). 2023-ban a világ lakosságának majdnem 60 százaléka, 4,76 milliárd ember használ valamilyen közösségi média platformot (Kemp, 2023). Az így előálló tartalmak mennyisége hatalmas. A közösségi média felhasználói által önkéntesen közzétett bejegyzések pedig mind alkalmasak lehetnek arra, hogy az

emberek véleményéről, gondolkodásról, cselekvésről, attitűdjéről, értékeiről képet adjon (Evans és Acaves, 2016).

Ezek az információk folyamatosan termelődnek, így vissza lehet nyúlni régebbi adatokhoz, meg lehet figyelni folyamatokat akár percről percre, de az adatfelvétel hosszas folyamatát átugorva, akár azonnali adatokkal is lehet dolgozni (Csepeli, 2015). Mindemellett, arra is lehetőség nyílik, hogy akár kisebb csoportok vagy ritka események is vizsgálhatóak legyenek (Kmetty, 2018). Az így előálló tartalmak tehát korábban nem látott szélességét és mélységet kínálnak szociológiai elemzéseknek. Ugyanakkor, ez a hatalmas mennyiségű adat, amely rendelkezésre áll, újabb módszertani megoldásokat is igényel. Az így keletkező tartalmak egy része kép, videó, illetve hanganyag, nagy többsége azonban szöveg. Ez azt jelenti, hogy az online térben rendkívül sok és változatos szöveges információ áll rendelkezésre.

2.2. A szöveges adatelemzés helye a szociológiai módszertanokon belül

A társadalomtudományos kutatások módszertanai hagyományosan a kvantitatív, illetve a kvalitatív módszerek. Babbie (2017) és Hajdu (2018) az alábbiakkal jellemzik e két módszertant. A kvantitatív módszerek statisztikai, matematikai eszközökkel vizsgálják a kérdéseket. E módszertan esetében nagy mennyiségű adat bevonása is lehetséges. A statisztikai próbák segítségével pedig a vizsgált minta alapján egy nagyobb mintára, vagy a teljes populációkra is becsléseket lehet adni. E módszertan célja, az objektív, általánosítható eredmény. Emiatt a mérések megismételhetőek, ellenőrizhetőek. Ilyen módszerek például a kérdőívek, a kísérletek vagy a másodelemzések. A módszertant jellemző általánosítás egyik fő hátránya, hogy az eredmények nem elég részletesek. Ezzel szemben a kvalitatív módszertan előnye éppen a kinyerhető információ gazdagsága. Kvalitatív módszer például az interjú, a terepkutatás vagy a fókuszcsoportos interjú. Ezek eredményei nem általánosíthatóak és nem megismételhetőek, mégis alkalmasak egy szűkebb kutatási kérdés megválaszolásra. A kvalitatív módszerek tehát egy árnyaltabb képet adnak a vizsgált jelenségekről, értelemadásokról és azok megértésére törekednek.

Mindkét bemutatott módszertan alkalmazható a szövegelemzésnél. Azonban, az eltérő megközelítés miatt a szövegekből kinyert tartalom, valamint azok elemzése is eltérő. Katona (2023) megállapítása szerint a szövegelemzésnél a kinyert tartalom két csoportba

sorolható. Ezt elsősorban az elemzés során használt módszer befolyásolja. A csoportosítás szerint a tartalom típusa lehet manifeszt és látens. A tartalomelemzés kvantitatív módszerekkel elsősorban a manifeszt, tehát nyilvánvaló tartalmak kinyerésére alkalmas. Az ilyen elemzések során a szöveg egyes elemei, például a szavak vagy szókapcsolatok, az egységek. A kiindulópont ezek megszámolása, amely segítségével számszerűsített adatokkal lehet dolgozni. A kvantitatív elemzések az ilyen elemek előfordulási gyakoriságával a szövegben lévő mintázatok feltárására törekszik. A számszerűsítés miatt az eljárás során statisztikai módszerek alkalmazásával történhet az elemzés.

A Katona (2023) által összefoglaltakat folytatva, a tartalomelemzés kvalitatív módszerekkel a látens, tehát rejtett tartalmak kinyerésére alkalmazható. „Segítségével a szövegeket kommunikációs kontextusukban elemezhetjük, bennük mintázatokat kereshetünk, és az adatokat klasszifikálhatjuk – anélkül, hogy szövegeinket kvantifikálnánk" (Katona, 2023, p.77). Az ilyen mintázatokat a szöveg egyes elemeinek kódolásával lehet kinyerni. Ezek alapján olyan leírások, beszámolók jönnek létre, melyek nem objektívak (Zhang és Wildemouth, 2017).

A két módszertan korlátainak átlépése, ezáltal pedig a jelenségek komplexebb leírása is lehetséges a kevert módszertannal (Király és mtsai, 2014). Ez ugyanis e két, kvantitatív és kvalitatív módszert ötvözi. A több módszertant használó kutatás ettől eltér. Ez utóbbin a hangsúly az egyik módszertanon van, és azt egészíti ki a másik. A kevert módszertan használatakor azonban a módszerek használata összhangban van. A gazdag információtartalom és feltáró jelleg mellett az eredmények számszerűsíthetése és általánosíthatósága is teljesül (Nagy Hesse-Bieber, 2010). Király és szerzőtársai (2014) megjegyzik, hogy mindez alapvetése, hogy a különböző minőségű adatokat lehetséges együtt kezelni. Tehát, a „kvantitatív adatok értelmezését segíthetik a kvalitatív vizsgálatok, míg a kvalitatív eredmények ellenőrzésében és validálásában a kvantitatív módszereknek lehet kiemelt szerepe" (Király és mtsai, 2014, p.96). A kevert módszertan tehát az eltérő megközelítéseket egy elméleti keretrendszeren belül, egymással összekapcsolva alkalmazza (Creswell és Plano Clark, 2007). Ahogyan a két módszertan külön-külön, úgy e harmadik, kevert módszertani megközelítés is alkalmazható szövegelemzéshez.

2.3. Az automatizált szövegelemzés

A korábban alkalmazott manuális szövegelemzés az információs társadalmat jellemző, nagy mennyiségű szöveges tartalom elemzésekor már egyre kevésbé alkalmazható. Az automatizált szövegelemzés azonban olyan mennyiségű szöveges tartalom elemzését teszi lehetővé rövid időn belül, ami az ember számára kivitelezhetetlen lenne.

Az automatizált szövegelemzés a természetes nyelvfeldolgozás egyik eszköze, amely hatalmas mennyiségű strukturálatlan szövegből képes az információ kinyerésére, rendszerezésére, megértésére, konceptualizálására és felhasználására (King, é.n.). Egyik problematikája tehát, hogy strukturálatlan adatokból dolgozik. Az ilyen adatokkal való munka nehézségét, ahogy az a nevében is benne van, annak rendezetlensége okozza, ugyanis az adatok nincsenek sorokba, oszlopokba rendezve (Németh és mtsai, 2020). A szövegelemzés egyik fontos feltétele az előfeldolgozás, amely során az adatok több lépésben kerülnek olyan formába, amellyel dolgozni lehet. Az előfeldolgozás lépései személyre szabhatóak, sorrendjük változhat. Ebből következik, hogy az előfeldolgozás egy olyan kritikus lépés, amely befolyásolhatja az elemzés kimenetelét is (Chai, 2023).

SZÖVEG	TOKENIZÁLÁS	KISBETŰSÍTÉS	LEMMATIZÁLÁS	STOPSZAVAZÁS
"Kivándorlás az országból"	Kivándorlás az országból	kivándorlás az országból	kivándorlás az ország	kivándorlás ország

Ábra 1. Az előfeldolgozás lépései. Forrás: saját szerkesztés

Az előfeldolgozás egyik első lépése a tokenizálás, ami a szöveg egységeire bontását jelenti. Az ilyen egységeket általában szóköz választja el egymástól. Ezen esetekben a szavak az egyes tokenek. Az 1. ábrán látható szöveg három szóból áll, így a tokenizálás után három egységre bomlik. Tokenek, tehát egységek azonban nem csak az egyes szavak lehetnek, hanem szavak részei vagy akár több szó is (Grefenstette, 1999). Az előfeldolgozás egy másik lépése a kisbetűsítés, ami az előforduló összes nagybetű kisbetűs változatára való lecserélését jelenti. Ez többek között a mondatkezdő nagybetűk miatt fontos. A „Kivándorlás” és a „kivándorlás” szavak ugyanis, az automatizált szövegelemzés során, a mondatkezdő nagybetű miatt eltérnek egymástól. A szöveg összes szavának kisbetűsítése azonban megoldja ezt a problémát. Egy

másik lépés, a lemmatizálás a szavak szótövesítését jelenti. Ez segít abban, hogy az azonos jelentésű szavak a ragozás miatt ne térjenek el egymástól az elemzés során. Az 1. ábra példáján, az „országból” szóról a „-ból” toldalék lekerül és csak az „ország”, az eredeti szó szótári alakja marad meg. Egy másik fontos lépés a stopszavazás vagy más szóval tiltószavazás, amely során kiválasztásra kerülnek azon szavak, melyek nem fogják az elemzés részét képezni. Az ilyen szavak kutatásonként változnak, hiszen ez egyes kutatási témáknál más szavak lesznek feleslegesek. Azonban, a stopszavazás során gyakran eltávolításra kerülnek a névelők, az írásjelek, illetve a létige.

Az előfeldolgozás során a lemmatizálás elsősorban a szózsák-modellt (bag of words) alkalmazó eszközöknél fontos. A szózsák-modell egy halmazaként kezeli a szövegeket. A szövegek egységekre bontva, szavakként szerepelnek a modellben. A szavak szövegbeli sorrendje, szintaktikai hierarchiája nincs figyelembe véve (Németh és mtsai, 2020, p.50). Ez azt jelenti, hogy e modellben a szavak a mondatban, illetve kontextusban betöltött szerepük, jelentésük alapján nincsenek értelmezve. Az ebből adódó problémák a következő példával szemléltethetőek. Egy szöveg olvasásakor a kontextusból kiderül, hogy a „bot” szó egy fa pálcára, horgászbotra, ütésre, automatizált feladatokat végrehajtó szoftverre, vagy a robot rövidítésére utal. A szózsák-modellben azonban e jelentések összemosódnak, és az elemzés végén, a kutatón múlik, hogyan értelmezi a „bot” szót.

A szöveges adatok strukturálatlanságából adódó problémák és az előfeldolgozás, valamint az alkalmazott szövegelemző modell miatti hibalehetőségek mellett, az automatizált szövegelemzés egyik jelentős problémája, hogy komplex tudást igényel. A különböző nemzetek nyelveit is magában foglaló természetes nyelvek feldolgozása „az informatika, a mesterségesintelligencia-kutatás és a nyelvészet határterülete” (Németh és mtsai, 2020, p.6). Ebből is következik, hogy az automatizált szövegelemzéshez is olyan összetett tudás kell, ami nem mindenki számára elsajátítható. Elsősorban a programozói ismeret teszi magassá a belépési küszöböt.

2.4. A low-code és no-code eszközök

A korábbi manuális szövegelemzést ma már felválthatja, illetve kiegészítheti az automatizált szövegelemzés. A módszer fejlődésével ugyanis olyan különböző automatizált szövegelemzési eszközök is készültek, melyek kódolást csak kevéssé (low-code), vagy

egyáltalán nem (no-code) igényelnek. Ezek az eszközök mégis elősegítik a szövegek automatizált feldolgozását és elemzését. Alkalmazásuk és megértésük mindenki számára egyszerűen elsajátítható. Ebből adódik, hogy lehetőséget nyújt olyan szakemberek és hallgatók számára is, akik csak felhasználói szintű számítógépes ismeretekkel rendelkeznek. Ezen eszközök jellemzője, hogy olyan felhasználói felülettel rendelkeznek, mely elősegíti az automatizálást azok számára, akik nem, vagy csak kevés programozói tudással rendelkeznek (Silva és mtsai, 2023).

Ezek az eszközök funkcióikat, céljukat, alkalmazási lehetőségeiket és elérhetőségüket tekintve változatosak. Vannak online elérhető eszközök, például a Google Trends, de több eszköz szoftver formájában, tehát letöltés és telepítés után érhető el. Ez utóbbira példa a dolgozatban is használt Meaning Extraction Helper, valamint AntConc. Silva és szerzőtársai (2023) összefoglalója alapján a low-code és a no-code eszközök pozitívumai elsősorban a könnyű kezelés, az erőforrás-megtakarítás és a gyors prototipizálás. Ez utóbbi azt jelenti, hogy az ilyen eszközök használatakor hamar kiderül az alkalmazott módszer hatékonysága, megfelelősége.

Egy no-code eszköz például a Google Trends, amely a Google keresések számáról készít statisztikákat, azok földrajzi elhelyezkedését és nyelvét is figyelembe véve. Az eszköz egyszerű, azonban a Google nagy elérése miatt hatékonyan képes a laikus nyilvánosság aktuális, illetve múltbeli érdeklődésének vizsgálatára. Az elmúlt években számos területen, például az informatika vagy a gazdaságtan területén alkalmazták sikeresen különböző kutatásokhoz (Jun és mtsai, 2018). Mindemellett előrejelzésekhez is alkalmazható. Ginsberg és szerzőtársai (2009) például a Google Trends segítségével hamarabb megjósolták az influenza terjedését, mint az amerikai Betegségellenőrzési és Felügyelő Központ, amely az ország lakosságának egészségéért felel. A Google egy másik eszköze, az Ngram Viewer az angolszász könyvek szövegének elemzésével képes az nyelvhasználati, kulturális, illetve társadalmi változások követésére (Ophir, 2016).

A no-code eszközök egy másik funkciója lehet a szentiment-és emócióelemzés, amely a szövegben megjelenő érzelmeket detektálja (Szigeti, 2022). A pszichológia területén ehhez az LIWC egy gyakran használt eszköz. Ezen eszköz segítségével a kutatók hatékonyan tudták vizsgálni többek között a szeptember 11-ei terrortámadások okozta nyelvi változásokat (Cohn és mtsai, 2004) és a koronavírus időszakában a mentális egészséget (Monzani és mtsai, 2021).

A programozást igénylő szövegelemzési modellek fő tulajdonságai közé tartozik, hogy előfeldolgozást igényelnek. Az előfeldoglozás egyes lépései a dolgozat előző részében már bemutatásra kerültek. A programozást igénylő szövegelemzési modellek egyik legnagyobb előnye, hogy ezek a lépések a kutatáshoz igazíthatóak. Így például a kutató döntésén múlik, hogy mi képezze a kutatás egységét és melyek az elhanyagolható szavak. Ehhez, az egyik előfeldolgozási lépés, a stopszavazás során egy olyan lista összeállítása szükséges, amely tartalmazza azon szavakat, melyek nem képezik az elemzés részét. A listán szereplő szavak annyira kevés információt tartalmaznak, hogy nincs rájuk szükség az elemzésben, jelenlétük csak zajként szolgálna (Kaur és Buttar, 2018). Mindemellett, a modell más paraméterei is változtathatóak, amennyiben az illeszkedés nem megfelelő. Tehát, a programozással együtt jár a rugalmasság is. A személyre szabhatóság ugyanis, amellett, hogy sok időt és energiát, valamint programozói tudást igényel, mégis döntő fontosságú az elemzés végkimenetelének érvényessége és megbízhatósága szempontjából (Chai, 2023).

A low-code és a no-code eszközöknél a könnyű használat éppen e rugalmasság rovására mehet. Az ilyen eszközöknél fontos, hogy az egyszerűség minél kevesebb olyan hátránnyal járjon, amely a flexibilis lehetőségek feladását jelenti. Tehát, amellett, hogy nem, vagy csak kevés programozást igényel az eszköz, fontos a beállítási lehetőségek megléte is. Ilyen például a Stopszó Lista megléte, szerkeszthetősége, a lemmatizálás megléte, szabályozhatósága, a kisbetűsítés lehetősége, illetve a lehetőség az elemzés egységének megadására. E dolgozatban két, az automatizált szövegelemzés kódolást nem igénylő eszköze kerül bemutatásra. Ezen eszközöket e feljebb megadott, személyre szabhatóság szempontjából értékelem. Az eszközök a kivándorlással kapcsolatos diskurzusok elemzése példáján kerülnek bemutatásra.

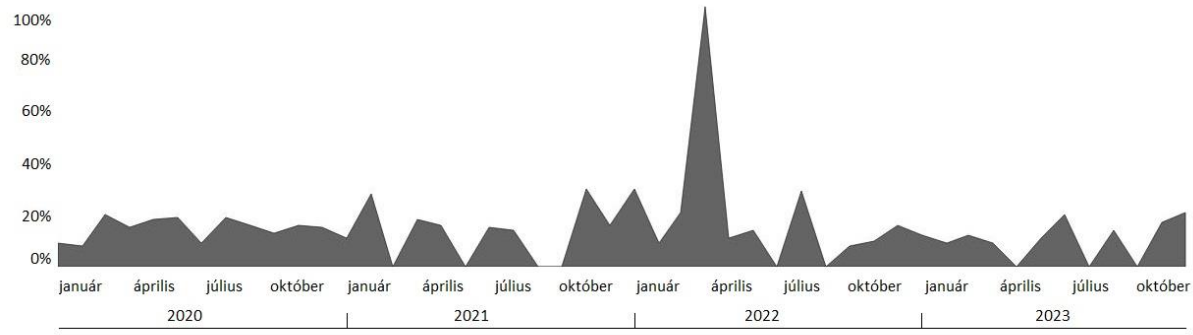
2.5 A kivándorlás

Az Európai Unió lakossága, bár egyre kisebb mértékben, de növekedést mutat a népességszám szempontjából. Magyarországon ez a tendencia már a nyolcvanas évek elején megfordult és folyamatosan csökken (Központi Statisztikai Hivatal, é.n.[a]). E létszámcsökkenés egyik oka a természetes fogyás, amely azt jelenti, hogy a halálozások száma magasabb a születések számához képest. Ezzel szorosan összefügg az idősödés, ami a népesség nem és kor szerinti összetételét ábrázoló korfákon is szembetűnő. A születések alacsony számára több magyarázat is lehet. A KSH Népességtudományi Kutatóintézete felmérése szerint a

gyermekvállalási tervek beteljesületlenségének elsősorban családi, egészségügyi és anyagi okai vannak (Kapitány, 2012).

A születési és halálozási számok mellett a vándorlás is fontos összetevője a népességszám alakulásának. A Magyarországra történő bevándorlás és annak problematikája az európai menekültválság óta sokakat foglalkoztat. A politikában és közbeszédben betöltött kiemelkedő szerepe mellett számos ezzel foglalkozó tanulmány is született az elmúlt évek során (Simonovits, 2016; Barna és Koltai, 2019). A vándorlás egy másik összetevője a visszavándorlás. A külföldről hazatérő magyarok száma nehezen mérhető, ezért pontos adatok nem állnak rendelkezésre, az azonban megfigyelhető, hogy az utóbbi években növekedést mutatott (Gödri és Horváth, 2021). E szakdolgozat azonban a vándorlás egy harmadik, kevéssé vizsgált összetevőjével, a kivándorlással foglalkozik.

A 2022-es magyarországi országgyűlési választások után kiemelt figyelmet kapott a Magyarországról történő kivándorlás is az általános közbeszédben. Több hírportál közölt ezzel foglalkozó írásokat, videó riportokat és a „kivándorlás" szóra való internetes keresések száma is kiugróan magas lett, ahogy az a 2. ábrán is látható.



Ábra 2. Google keresések a „kivándorlás" szóra 2020.01.01 és 2022.12.31. között. A számok a keresési érdeklődést jelzik a grafikon legmagasabb pontjához képest Magyarországon. Forrás: saját szerkesztés a Google Trends adatai alapján

A Központi Statisztikai Hivatal (é.n.[b]) adataiból később az is kiderült, hogy a kivándorlás nem csak a közbeszédben volt jelen, hanem az a gyakorlatban is megmutatkozott. A Magyarországról kivándorlók száma, míg 2015-től folyamatosan csökkenő tendenciát mutatott, egy azt megelőző évben mérsékelt növekedést követően 2022-ben jelentősen megnőtt. Bíró-Nagy és Szabó (2021) kutatása rámutat, hogy e téma miért ekkor kerülhetett központba. Azt találták, hogy „Magyarországon a kivándorlási hajlandóság tekintetében a legjelentősebb különbségek a politikai csoportok között vannak" (p.35). „Háromszor annyi

ellenzéki (36%) és kétszer annyi bizonytalan szavazó (24%) szeretné elhagyni az országot, mint kormánypárti fiatal (12%)" (p.36). Siskáné Szilasi és szerzőtársai (2017) szintén kiemelték, hogy a magyar fiatalok erősödő kivándorlási szándékának a „politikai élettel való növekvő elégedetlenség, az életkilátások beszűkülése, a bizonytalanság miatti jövővel kapcsolatos félelmek és a társadalmi konfliktusok kiéleződése" (p.145) a fő okai. Mindezt kiegészítik olyan úgynevezett vonzó és taszító tényezők (Ravenstein, 1885), melyek a kivándorlót a hazájától taszítják és a fogadó ország felé vonzzák. Az ilyen tényezők elsősorban gazdasági jellegűek (Hárs, 2020; Horváth, 2023). Így például a fogadó országban lévő magasabb fizetések az ország felé vonzzák, míg az otthoni munkalehetőségek hiánya Magyarországról taszítják a kivándorlókat.

2.6. A kivándorlás-diskurzusok vizsgálata

A fent ismertetett kutatások interjúkkal, illetve kérdőívekkel készültek. Ezek hátrányai, hogy nem beavatkozásmentesek, tehát nagyon sok tényező, például az interjúztató személye, a környezeti tényezők, vagy a kérdések megfogalmazása befolyásolhatja az eredményeket is (Katona, 2023). A kivándorlási szándék okai, az azzal kapcsolatos vélekedések azonban a fent ismertetett kutatások módszerei mellett további módszerekkel is vizsgálhatóak. A közösségi média diskurzusok elemzése lehetőséget kínál arra, hogy mélyebb betekintést nyerjünk az emberek e témával kapcsolatos véleményeibe, amelyre a hagyományos kérdőíves felmérések nem feltétlenül alkalmasak. Dessewffy és Láng (2015) kiemeli, hogy a kérdőíves felmérések egyik hátránya, hogy csak előre meghatározott kérdésekkel kapcsolatos attitűdöket tudnak mérni. Ilyen esetekben a válaszadóknak „legtöbbször nincs kikristályosodott és tiszta álláspontunk az adott kérdésekről" (p.163), emiatt pedig a válaszuk is „homályos és gyakran nem létező nézeteket tükröz" (p.163). A közösségi médiában való megnyilvánulás ezzel szemben önkéntes. A felhasználó közölni szeretné véleményét másokkal, saját maga által választott témákban. Az így leírtak tehát organikusan jönnek létre.

E dolgozatban a hazai laikus közvélemény kivándorlással kapcsolatos diskurzusait fogom vizsgálni és az abban megjelenő fő témákat vetem össze az online sajtó kivándorlással kapcsolatos cikkeinek fő témáival. E dolgozatban a diskurzus kifejezést a laikus közvélemény, illetve a sajtó által közzétett szövegeket foglalja magában. A „diskurzuselemzés során természetes módon létrejövő szövegeket vizsgálnak, melyeknek fontos a társadalmi

kontextusa, a szereplők viszonyai, és azok a szerepek, amelyekben megszólalnak” (Katona és mtsai, 2021, p.76).

Azt, hogy milyen diskurzusok jelentek meg a Magyarországról történő kivándorlásról a laikus közvéleményben, illetve az online sajtóban, korábbi kutatásokban még nem vizsgálták. Ugyanakkor, a laikus közvélemény bizonyos platformjait már elemezték a kivándorlással kapcsolatban. Görög (2017) például tíz vlog elemzésén keresztül mutatta be, hogy a kivándorolt magyarok hogyan mutatják be külföldi tartózkodásukat és milyen lehetőségekhez jutnak ez által. Dolgozatom azonban az olyan bejegyzéseket és hozzászólásokat elemzi, amelyek szerzői nem feltétlenül külföldön élő magyarok. Így e dolgozat a kivándorlással kapcsolatos laikus közvélemény diskurzusainak szélesebb körű bemutatására törekszik. E törekvés gyakorlati megvalósításához elengedhetetlen a laikus közvélemény, illetve a sajtóorgánumok által közzétett nagy mennyiségű hozzászólás, bejegyzés és cikk elemzése. Ezek gyors és hatékony feldolgozásához segítséget nyújtanak az automatizált szövegelemzési eljárások és eszközök.

3. A használt eszközök

3.1. A Meaning Extarction Helper bemutatása

3.1.1 Az eljárás és az előfeldolgozás lépései

A Meaning Extraction Method (a továbbiakban: MEM) egy olyan szövegelemzési eljárás, amely során főkompones-elemzést alkalmazva csoportosíthatóak egy korpuszban megjelenő szavak és megnevezhetőek az abban megjelenő témák. Az MEM tehát egy olyan keretrendszer, amely a korpusz témáinak beazonosítására alkalmas. Ezt az eljárást Chung és Pennebaker (2008) fejlesztette ki és használta arra a célra, hogy az emberek saját magukról alkotott képét vizsgálják. A szerzők már ezen írásukban kifejtették, hogy a kidolgozott eljárás nem csak a pszichológia, hanem más tudományokágak kérdéseinek megválaszolására is alkalmazható. Boyd (2017) leírása alapján az MEM eljárás három alapvető lépésből áll. Első lépés a korpuszban gyakran előforduló szavak beazonosítása. Második lépés egy olyan táblázat létrehozása, amely tartalmazza, hogy az egyes szövegek mely gyakori szavakat tartalmazzák. A harmadik lépés, azon szavak csoportosítása statisztikai módszerekkel, melyek együtt fordulnak elő a korpuszban.

A Meaning Extraction Helper (a továbbiakban: MEH) (Boyd, 2022) egy olyan automatizált szövegelemzési eszköz, amely az MEM eljáráshoz nyújt segítséget. Az eszköz használatakor a korpusz előfeldolgozásának lépései állíthatóak be. Az eszköz a korpuszt alkotó összes szöveget tokenizálja, kisbetűsíti, lemmatizálja, kiszűri a tiltószavakat, a ritkán használt szavakat és a túl rövid szövegeket, valamint beállítja, hogy maximum hány szóból álló egységekre terjedjen ki az elemzés. Ezt követően az eszköz kimeneti fájlokat hoz létre, amelyek az MEM eljárás első két lépését teljesítik, de akár az eljárástól függetlenül, önmagukban is értelmezhetőek. A harmadik lépéshez egy statisztikai program szükséges. Az eszköz használatával végezhető előfeldolgozás az alábbiak alapján történik.

A **Token Handling Options menüpont**ban több beállítás található. A tokenizálás a szöveg egységeire bontását jelenti. Az MEH az egységeket szóköz alapján (a programban: Whitespace Tokenizer) is tudja tagolni, de a közösségi médiában közzétett bejegyzéseket érdemesebb a Twitter-Aware Tokenizer-t használni. Ez utóbbi például az olyan plusz jelentéssel bíró írásjeleket, mint az „@” és a „#” külön kezeli. A menüpontban továbbá beállítható a kisbetűsítés is. A kisbetűsítés a korpuszban előforduló összes nagybetű kisbetűs változatára való lecserélését jelenti. Ez azért fontos, mert az eszköz a kis-és nagybetűket külön kezeli. Emiatt, a kisbetűsítés nélkül egy, a mondat első helyén álló szó külön szónak számítana a kis kezdőbetűvel írt, de ugyanazon betűket, megegyező sorrendben tartalmazó szótól. További pozitívuma, hogy a közösségi médiában közzétett posztokban a kis-és nagybetűk használata nem mindig felel meg a helyesírási szabályoknak, tehát nem egységes. A felhasználók többek között a földrajzi neveket és az azokból létrehozott mellékneveket sokszor helytelenül írják („az egész magyarország neki legyen rabszolgája”; „ne felejtsük el a Magyar megélhetési kivándorlást se”) vagy nagybetűkkel jelölik egy szó kiemelt jelentőségét („Na akkor mégegyszer ezek TÉNYEK!”). A kisbetűsítésben hibalehetőség is rejlik, ugyanis vannak szavak, melyeknél a nagy, valamint a kis kezdőbetűvel írt változatok jelentése eltér. A Virág személynév és a virág főnév például külön jelentéssel bírnak, a kisbetűsítés miatt azonban összeolvadnak. A dolgozat korpusza miatt azonban e hibalehetőség elhanyagolható.

A **Conversation List menüpont**ban egy másik fontos lépés, a lemmatizálás található. Ez a szavak szótövezését jelenti. Ennek segítségével a program csak a szavak szótövét vizsgálja, ezzel elkerülve azt a hibát, hogy egy szó azonos jelentéssel többször is szerepeljen a kimeneti listán. Példaként, a lemmatizálás során mind az „államoknak”, az „államon”, az „államot” és

az „államtól” szavak átváltásra kerülnek és az „állam” szóként kerülnek be az elemzésbe. Fontos megjegyezni azonban, hogy a MEH által automatikusan létrehozott Átváltási Lista nem kimerítő, az elemzés témájának megfelelően előfordulhatnak olyan szavak, melyek nem szerepelnek az Átváltási Listán. Ez azt is jelenti, hogy a lemmatizálás nem nyelvtani szabályok, hanem egy lista alapján történik. Ez az Átváltási Lista azonban egyszerűen lehet szerkeszteni, módosítani. Fontos továbbá, hogy a korpusz nyelve e menüpontban beállításra kerüljön. A program listájából kiválasztható 25 nyelv a magyar nyelvet is tartalmazza.

A **Stop List menüpont**ban a korpusz nyelvének újbóli kiválasztása után létrejön egy nyelvspecifikus, szerkeszthető tiltószó lista. E listára azon írásjelek, különleges karakterek (például idézőjel, csillag), számok és bizonyos szavak (például névelők, kötőszavak, létige) kerülnek, melyek nem fogják az elemzés részét képezni. Fontos, hogy ezek a szavak ne képezzék az elemzés részét, ugyanis nem bírnak tartalmi jelentéssel. Azonban fontos megjegyezni, hogy mindig az adott kutatás függvénye, hogy mely szavak nem bírnak jelentőséggel, ezért fontos, hogy e lista a kutatók által változtatható legyen. Az MEH tehát, ha a tiltószó lista egyik elemét találja a szövegben, annak előfordulását nem számolja és nem jeleníti meg a kimeneti fájlokban.

Emellett, az **N-gram Settings menüpont**ban további szavakat és szövegeket is ki lehet szűrni. Az utóbbi a szöveget alkotó szavak száma alapján lehetséges: meg lehet adni, hogy mi legyen az egyes szövegek minimum szószáma. Ez azt jelenti, hogy a megadottnál rövidebb szövegek nem kerülnek be az elemzésbe. Emellett a ritkán előforduló szavakat is ki lehet szűrni az elemzésből, több szempont alapján. Ritkán előforduló szónak számíthatnak azon szavak, melyek a szövegek vagy a korpusz egészének egy bizonyos százalékában nem szerepelnek. Mindemellett beállításra kerülhet, hogy ne csak az egyes szavak, hanem megadott szóból álló kifejezések előfordulását is vizsgálja a program. Az MEH ennek paramétereit A és B betűkkel jelöli. Ha például A értéke 2, B értéke 4, akkor a program az összes 2, 3 vagy 4 szóból álló szócsoportok előfordulását vizsgálja. Ezeket a szó-n-eseteket nevezi az angol terminológia n-gramnak, innen ered a menüpont elnevezése.

3.1.2. A kimeneti fájlok

Az MEH eszköz a korpusz előfeldolgozása után két olyan kimeneti fájlt generál, amelyet tovább lehet elemezni: a Gyakorisági Listát (Frequency List) és a dokumentum-kifejezés

mátrixot (Document Term Matrix; a programban: DTM Binary). A Gyakorisági Lista egy olyan, akár több ezer sorból álló lista, melyen az eszköz által elemzett szavak szerepelnek. Az összes szóhoz tartoznak különböző értékek is, melyek mind annak előfordulására vonatkoznak. Az első ilyen érték maga a gyakoriság, ami azt mutatja, hogy a korpuszban összesen hányszor található meg az adott szó. Ennek sorrendbe rendezése után megfigyelhető, hogy mely szavak szerepelnek a legtöbbet a korpusz egészében. Emellett, a Gyakorisági Lista tartalmazza, hogy hány dokumentumban szerepel az adott szó, az összes dokumentum hány százalékában szerepel az adott szó, valamint, hogy mennyire egyedi az adott szó[1]. A Gyakorisági Listára kerülő szavak különböző kritériumok szerint, előfordulásuk alapján beállításra kerülhetnek az N-gram Settings menüpontban.

	szó1	szó2	szó3	szó4	szó5	...
dokumentum1	1	1	1	1	1	
dokumentum2	0	0	0	1	0	
dokumentum3	0	0	0	0	0	
dokumentum4	1	0	0	0	0	
dokumentum5	0	0	0	0	0	
...						

Ábra 3. A dokumentum-kifejezés mátrix felépítése. Forrás: saját szerkesztés

Az eszköz által létrejön továbbá a dokumentum-kifejezés mátrix, amely egy olyan mátrix, ahol a sorokban a korpusz egyes dokumentumai (megfigyelések), míg az oszlopokban a Gyakorisági Listában is megtalálható szavak (változók) találhatóak. Amennyiben az adott dokumentumban szerepel az adott szó, a cella értéke 1 lesz, ha nem szerepel benne, a cella értéke 0 lesz. Az 3. ábrán például az összes látható szó szerepel az első dokumentumban, míg az ötödik dokumentumban egyik sem. A negyedik szó az egyetlen, amely mind az első, mind a második dokumentumban előfordul. A dokumentum-kifejezés mátrixban tehát a szavak előfordulásának gyakorisága nem számít. Az MEM eljárás csak a 0 és 1 értékeket használatával a természetes nyelvek sajátosságához igazodik. „Természetes nyelvnek azokat a nyelveket nevezzük, amelyek spontán alakultak ki, amelyek nyelvtana az emberek közötti nyelvi

[1] Ezt az inverz dokumentumgyakoriság (a programban: IDF) mutatja, amely a korpuszban található dokumentumok számának logaritmusa, osztva azon dokumentumok számával, amelyekben az adott kifejezés szerepel. A IDF annál kisebb, minél több dokumentumban szerepel az adott szó. Így például a névelők nagyon kis súlyt kapnak. (forrás: https://tfidf.com)

kommunikáció természetes fejlődésének eredményeként alakult ki" (Németh és mtsai, 2020, p.6). Az ilyen nyelvek nyelvtana generatív (Chomsky, 1957), tehát véges számú nyelvi elemből végtelen számú mondatot lehet alkotni, például szinonimák használatával. Az egyes szavak leggyakoribb értéke így a 0 (Chung és Pennebaker, 2008), hiszen a beszédben sok alternatív szó használható helyette.

3.1.3. Dimenziócsökkentés

A dokumentum-kifejezés mátrix több ezer sorból és oszlopból is állhat, emiatt önmagában nehezen érthető. Dimenziócsökkentésre van szükség, hogy könnyebben lehessen értelmezni az adatokat. A társadalomtudományokban a dimenziócsökkentéshez a főkomponens-elemzés (Principal Component Analysis, a továbbiakban PCA) az egyik leggyakrabban használt megközelítés (Markowitz, 2021). A klasszikus társadalomtudományos kutatások a PCA-t elsősorban közvetlenül nem mérhető attitűdök mérésére használják. Az olyan attitűd például, mint az idegenellenesség, nem feltétlenül derül ki, ha erre közvetlenül rákérdeznek. Több más, ehhez kapcsolódó kérdés azonban választ adhat arra is, hogy a kérdezett elutasító-e az idegenekkel szemben. A PCA során ezen többi manifeszt változó segítségével jön létre egy látens, tehát rejtett változó, ami az idegenellenességet mutatja. Ha több látens dimenzió van, az a manifeszt itemek egy-egy csoportjaként határozható meg, ahol a főkomponens és az item korrelációja, illetve annak előjele ad támpontot ehhez.

Az MEM eljárás azonban ettől eltérően alkalmazza a PCA-t. A dokumentum-kifejezés mátrixban szereplő szavak az egyes változók, amelyekből nagy információval rendelkező látens dimenziók jönnek létre. Erre úgy is lehet tekinteni, hogy a szavak a PCA után csoportokat képeznek, amelyek így egy-egy témát jelölnek. A szavak ilyen módszerrel való csoportosítása azon az elgondoláson alapszik, hogy az együtt előforduló szavak pszichológiailag értelmes szócsoportokhoz vezethetnek (Chung és Pennebaker, 2008). Ha például két szó gyakran együtt fordul elő, korrelál, akkor valószínű, hogy valamilyen szemantikai kapcsolat áll fenn közöttük. Az MEM során alkalmazott dimenziócsökkentés részben a Markowitz (2021) tanulmányában leírtakkal magyarázom a következő sorokban. A szavak alkotta csoportok az egyes főkomponensek, melyek az alkalmazott varimax rotáció miatt függetlenek egymástól. Az egy szócsoportot alkotó szavak faktorsúlya a PCA modell által adott faktorsúlyhoz hasonló, és a többi szócsoportot alkotó szó faktorsúlyától eltérő. Ezek a faktorsúlyok általában kisebbek a

hagyományos kérdőívekéhez képest, amely okai szintén a természetes nyelv fentebb említett sajátosságaiban, generatív nyelvtanában gyökerezik.

3.1.4. Korábbi alkalmazások

Az MEH eszközt számos terület különböző kérdéseinek vizsgálatára alkalmazták az elmúlt évek során, többek között a média (Rhidenour és mtsai, 2019), az aktivizmus (Pham és mtsai, 2023), az értékek (Wilson és mtsai, 2016), a kulturális különbségek (Rodríguez-Arauz és mtsai, 2017), valamint a mentális egészség (Currin-McCulloch és mtsai, 2020) vizsgálatában. Ezen kutatások az alábbiakban bemutatásra is kerülnek. Közös bennük, hogy az MEH eszközt meglévő korpuszok szógyakoriság listájának kinyerésére alkalmazzák, valamint az eszköz által létrehozott dokumentum-kifejezés mátrixot használva PCA-t alkalmaznak egy statisztikai program, általában az SPSS segítségével. Az elemzésekben az egyes szavakat vizsgálják és nem a több szóból álló szókapcsolatokat.

Rhidenour és szerzőtársai (2019) azt vizsgálták, hogy Amerikában, a különböző években hogyan ábrázolja az elit hírmédia a veteránokat a Veteránok Napján és annak környékén. Az MEH használatával a kutatók az újságcikkekből álló korpuszból kinyerték a leggyakrabban előforduló szavak listáját. Ezt követően a dokumentum-kifejezés mátrixon végzett PCA által kinyert faktorokat - tehát a korpuszban előforduló szavak alkotta csoportokat - tematikusan felcímkézték. A faktorokat három kutató egymástól függetlenül címkézte fel, korábbi irodalmakban megjelenő veterán ábrázolások alapján. Ha a különböző kutatók különböző eredményre jutottak egy-egy faktor tematikus elnevezésekor, az eredményeket összehasonlították, majd megismételték a folyamatot. Az így kapott címkék tekinthetőek a korpuszban előforduló témáknak. Minden téma esetében, a szerzők csoportosították, hogy a média áldozatként vagy hősként keretezi a veteránokat.

Pham és szerzőtársai (2023) a rasszizmus ellenes aktivisták céljait és motivációit vizsgálták. A kutatás korpuszát egy online kérdőívben szereplő két nyitott kérdésre adott válaszok alkották. A két kérdésre adott válaszokban előforduló szavak gyakoriságát külön vizsgálták, a szövegek kevesebb, mint 0,5 százalékában előforduló szavak kiszűrésével. Ez utóbbi határértéket a válaszok rövidségével magyarázták. Ezt követően a szerzők a PCA segítségével kapott szócsoportokat felcímkézték, így létrehozva a témák elnevezését. Ezeket

a témákat tekintették a laikus vélekedés szerinti rasszizmusellenes aktivizmus fő céljainak és motivációinak.

Wilson és szerzőtársai (2016) az egyéni értékeket, azok viselkedéssel való kapcsolatát és kimutatását vizsgálták amerikai és indiai résztvevőkkel. Ezen elemzés elsősorban a módszertan bemutatására összpontosított. Mind nyitott kérdésekre adott válaszokat, mind egy blogról legyűjtött bejegyzéseket korpuszként használtak elemzésükhöz. Az MEH elemzést felhasználva, a dokumentumok legalább 5 százalékában előforduló szavak elemzésével, szintén PCA-t alkalmazva nyerték ki a kérdőíves kérdésekre adott válaszokból létrehozott korpuszban az előforduló témákat. A blogbejegyzésekre külön nem alkalmazták az MEM-et.

Rodríguez-Arauz és szerzőtársai (2017) nyitott kérdésekre adott válaszok alapján figyelték meg azt, hogy az amerikai-spanyol kétnyelvű emberek a nyelvek közötti váltás során megváltoztatják a saját magukról alkotott képüket is. A szerzők a MEH eszközt külön használták a spanyol, valamint az angol nyelvű szövegek elemzésére. Azon szavak kerültek a szólista mátrixba, amelyek a korpusz legalább 3,5 százalékában szerepelte, így 233, illetve 144 szót vizsgáltak tovább. Az eredményekre a fentiekkel azonos módon jutottak.

Currin-McCulloch és szerzőtársai (2020) a mellrák túlélők mellrákról szerzett általános tapasztalataiban, valamint az információkeresési diskurzusban megjelenő témákat kutatták. Ehhez a Reddit közösségi média platformról gyűjtött bejegyzések, és e bejegyzések legalább 7,5 százalékában megjelenő szavakat elemezték. Az MEH és a dokumentum kifejezés-mátrixon végzett PCA által kinyert témákat egyeztették fókuszcsoportos eredményekkel. Az eredmények hasonlósága miatt arra jutottak, hogy az MEM alkalmas a gyakran előforduló témák feltérképezésére.

3.2. Az AntConc bemutatása

Az AntConc (Anthony, 2023) egy olyan automatizált korpuszelemző eszközkészlet (Anthony, 2005), amelynek számos funkciója van. Az MEH-tól eltérően az AntConc elsősorban kulcsszavas keresés alapján képes a korpusz elemzésére, valamint az előfeldolgozás lépései nem állíthatóak be. Az AntConc-ot jellemzően a nyelvészetben használják (De Felice és mtsai, 2018, Furkó, 2019), de többek között a média (Ross és Rivers, 2018), valamint a társadalomtudomány (Wang, 2022) területén is alkalmazták már.

Furkó (2019) A kormánypárti és ellenzéki képviselők által tartott parlamenti beszédek nyelvhasználatának elemzésével a magyarországi populista diszkurzív stratégiákat vizsgálta. Ezen elemzésben, mások mellett az AnctConc **Keyness funkció**ját használta. E funkció használatához két korpusz szükséges, egy célkorpusz és egy referencia korpusz. Az ezekben megjelenő szavakat hasonlítja össze egymással és adja ki azon szavak listáját, amelyek előfordulása szignifikánsan jellemzőbb a cél-, mint a referencia korpuszban. Furkó a bemutatott tanulmányban mind a kormánypárti, mind az ellenzéki képviselők által tartott beszédeket referencia korpuszként használta a másik korpuszhoz képest. Az AntConc minden funkciója személyre szabható a Tool Settings menüpontban, ahol többek között a használt statisztikai tesztek és határértékek is kiválaszthatóak. Furkó statisztikai mérőszámként a Log Likelihood tesztet alkalmazta, ahol a 3,84-es Log Likelihood-érték volt a kritikus érték a szignifikánsan eltérő szógyakorsiág megítélésekor ($p<0,05$). Az így alkalmazott Keyness funkció segítségével megfigyelhető, hogy a kormánypárti beszédekben gyakoribbak a potenciálisan populista diszkurzív stratégiákkal társítható lexikai elemek.

Ross és Rivers (2018) Donald Trump Twitter bejegyzéseit vizsgálták. A Keyness funkciót Furkótól eltérő módon alkalmazták. A referencia korpuszt amerikai politikusok hitelesített felhasználói által közölt Twitter bejegyzésekből építették fel. Ezzel hasonlították össze a Trump által közölt bejegyzéseket, mint célkorpuszt. Ezzel a megoldással az amerikai elnök nyelvezetét vetették össze a többi politikuséval, mindkét esetben a Twiterre szorítva a korpuszt. Ez utóbbi azért fontos, mert eltérő platformok eltérő nyelvezetet kívánnak. Az így kinyert szavakat tovább elemezték a **Klaszter funkció**val (a programban: Cluster). Ez a funkció megjeleníti a keresett kifejezést körülvevő szócsoportokat (Anthony, 2005, p.11), tehát azokat a szavakat, amelyek a keresett szó vagy kifejezés közvetlen szomszédjai. A klaszterekből kiderült, hogy a vizsgált célkorpuszt alkotó Twitter bejegyzések kulcsszavainak többsége a médiával, valamint annak álhír terjesztő szerepével kapcsolatos. Mindemellett, az AntConc **Konkordancia funkció**ját is alkalmazták az elemzésben. A konkordancia (a programban: Keyword in Context, röviden: KWIC) a keresett szó közvetlen környezeti előfordulását mutatja meg. Ennek segítségével a szó kontextusba kerül, a program listába szedi az azt megelőző és azt követő szavakat, beállítás szerint. Ezen lista egészét vagy egy abból vett véletlen mintát is meg lehet jeleníteni. A lista sorrendjén beállításra kerülhet, hogy a keresett szó melletti szavak gyakoriság alapján legyenek rendezve, így megmutatva a leggyakoribb mintázatokat. Tehát, a

Twitter bejegyzések nem csak kulcsszavait, kifejezéseit, hanem azok kontextusát is vizsgálták a szerzők tanulmányukban. Mindebből arra a következtetésre jutottak, hogy Trump a Twittert a média hihetőségének megkérdőjelezésére és a saját szavahihetőségéről való meggyőzéshez használja.

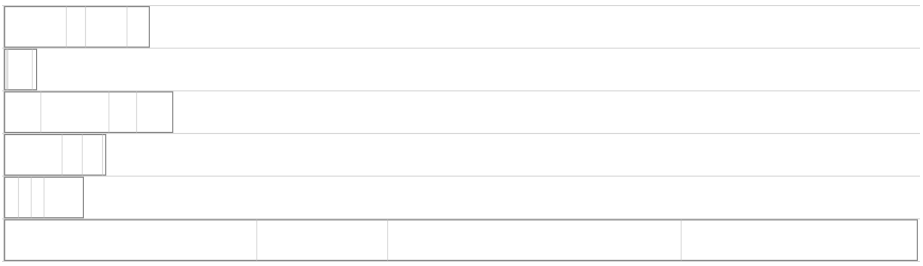
De Felice és szerzőtársai (2018) a státusz és a nem hatását vizsgálta az udvariassággal kapcsolatos nyelvi döntésekre. Korpuszukat az Egyesült Államok külügyminisztere által küldött, illetve az ő részére érkező nyilvánosságra hozott e-mail üzeneteket alkották. A korpuszt három részre osztották aszerint, hogy a feladó, illetve a címzett a Külügyminisztérium munkatársa vagy külsős személy. Az elemzéshez az AntConc **N-Gram funkció**ját használták. Az n-gram-ban n számú szó egy egységként kerül listázásra. N értéke beállítható, így, ha például n=3, akkor a leggyakrabban előforduló, a szövegben három egymás mellett lévő szó, mint egység kerül kilistázásra. E funkció mögött az a megfontolás áll, hogy nem az egyes szavak, hanem a szópárok alkotják a nyelvhasználatot jellemző elemzési egységet. Példaként, a „határon túli magyarok" egy 3-garm, az „európai unió" egy 2-gram. Az e-mailek elemzésénél n értéke 2 volt és az így kapott 2-gramok relatív gyakoriságát vizsgálták, kiegészítve ezen egységek konkordanciájával. Arra a következtetésre jutottak, hogy a hierarchia, valamint az udvariasság kevéssé jelenik meg az üzenetek nyelvhasználatában. Ehelyett ezek az e-mailek tartalmából és azok funkciójából olvasható ki.

Wang (2022) brit újságok által publikált cikkeket elemzett. Arra kereste a választ, hogy egy tüntetéssorozat kapcsán milyen szerepet töltöttek be fontosabb társadalmi szereplők. Az AntConc **Szólista funkció**jával (a programban: Word) kinyerte a korpuszban előforduló szavak listáját. E funkció az MEH szógyakoriság funkciójával hasonló eredményt hoz, amennyiben a szólista sorrendje gyakoriság szerint kerül beállításra. Fontos azonban, hogy az AntConc nem alkalmazza az MEH korpuszt előkészítő lépéseit, mint például a lemmatizálás vagy a stopszavazás. Wang e szólista alapján azonosította a cikkekben megjelenő társadalmi szereplőket. Bár a szerző ezt nem alkalmazta, a Szólista funkció a szavak listán belüli keresésre is alkalmas. Így listázni lehet az összes olyan korpuszban előforduló szót, amely egy bizonyos karakterrel, vagy karakterek sorozatával kezdődik. Wang tanulmányában a társadalmi szereplők beazonosítása után az AntConc **Kollokáció funkció**ját használta. E funkció a keresett kulcsszó közelében legnagyobb valószínűséggel előforduló szavakat mutatja. A tanulmányban a társadalmi szereplők voltak a kulcsszavak és a tőlük maximum 5 szó távolságra lévő és

legalább 5 alkalommal előforduló szavak kerültek az elemzésbe. Az így kapott eredményeket a szerző manuálisan három csoportra osztotta, melyek az adott társadalmi szereplők szerepét jelölték.

E fent ismertette funkciókat az AntConc használatával nem csak egymástól függetlenül lehet használni, ugyanis a menüpontok egymással interakcióban állnak. Például, a kollokáció során kapott egyes találatokra kattintva, megkapjuk azok közvetlen környezeti előfordulását, tehát konkordanciáját. Ezt követően, további kattintással a **Fájl nézet funkció** jelenik meg, ahol a kollokáció során kiválasztott szövegrészletet tartalmazó teljes szöveg tekinthető meg. Egy másik lehetőség továbbá, a Kollokáció funkció találatinak szófelhős ábrázolása **Wordcloud funkció**val.

Az eszköz egy korábban még nem ismertetett funkciója a **Plot funkció.** Ez a korpusz egyes szavainak vizualizálására alkalmas. A korpusz szövegei az egységek. Ezek jellennek egy horizontális oszlopgrafikon oszlopaiként (ábra 4.) A kulcsszavas keresés után, a szövegekben előforduló keresett szó egy függőleges vonallal kerül jelölésre az egyes oszlopokon. Beállításra kerülhet, hogy mindez normalizált formában legyen megjelenítve, ekkor az oszlopok egyenlő hosszúak. Egy nem normalizált beállításra kapott eredmény a 4. ábrán is látható. Az egyes oszlopok (szövegek) különböző hosszúságúak, a legalsó oszlop jeleníti meg a legtöbb szóból álló szöveget. Az átfedés (a programban: Overlay) opció bekapcsolásával, több szó egyszerre, különböző színű vonalakkal is ábrázolható. E funkciót korábbi kutatásokban nem használták.



Ábra 4. A „kiv?ndor kifejezés megjelenése a laikus közvélemény korpusz egyes szövegeiben. A sötétebb színnel határolt oszlopok az egyes szövegeket, azok hosszát ábrázolják. A függőleges vonalak a kifejezés helyét jelölik. Forrás: képernyőfotó az AntConc Plot funkciójának eredményéről, annak egy részlete*

4. A dolgozat elemzésének módszertana

4.1 Korpuszok

A korpusz a SentiOne közösségi figyelő eszköz (social listening tool) segítségével került kigyűjtésre, ahol egy megadott kulcsszó vagy kulcsszavak kombinációjával, valamint kizáró

szavak és keresési helyek megadásával lehet az interneten lévő, nyilvános hozzáférésű szövegeket legyűjteni. Két korpusz került így legyűjtésre. Mindkét korpusz összes szövegében szerepel egy „kivándor” szótővel kezdődő szó (pl „kivándorlás”, „kivándorolt”, „kivándorló”, stb). Emellett, az egyes szövegek forrása a közösségi média, illetve a hazai sajtó felületei, magyar nyelvű, valamint 2022. április 1. és 2022. szeptember 1. közötti szövegek. A vizsgált időintervallum az az évi magyar országgyűlési választások utáni időszak. A keresett időszak előtt kitört ukrán háború kapcsán, az onnan kivándorló emberekről szóló posztok miatt sok fals találat volt, melyek a keresőkifejezés módosításával[2] kiszűrésre kerültek. Mindkét korpusz adattisztításon ment keresztül. Első lépésben kiszűrésre kerültek az adatgyűjtésből adódó hibák: az olyan szövegek, amelyek azonosítója megegyező volt. Második lépésben kiszűrésre kerültek az ismétlődő elemek: az olyan szövegek, melyek tartalma teljesen megegyezett egymással. Ezek gyakori előfordulása a médiumok más médiumok által közölt cikkeinek magas számával magyarázható. Ez utóbbi döntés és annak eredményessége a no-code eszközök limitációi című alfejezetben bővebb kifejtésre kerül.

Az egyik korpuszt a laikus közvélemény által közzétett hozzászólások és bejegyzések alkotják. Forrását tekintve, ez az összes olyan online tartalmat magába foglalja, ami nem szerkesztett média. Az így kapott, körülbelül 37 000 találatból, a kezelhetőség miatt egy 20 százalékos egyszerű véletlen mintavétel következett. A korpusz a legyűjtés után 7 440 elemből, majd az adattisztítás után 7 439 (megegyező azonosító kiszűrése), illetve 7 298 elemből (ismétlődő elemek kiszűrése) állt. Ahogyan az az 5. ábrán is látható, forrását tekintve a korpusz egyenlőtlenül oszlik meg, közel 85 százalékát a Facebookon közzétett bejegyzések és hozzászólások teszik ki.

Facebook	Websites	Reddit	Blogs	Forums	X (Twitter)	Instagram	YouTube	Reviews
6 263	906	140	68	39	21	1	1	1
84,18%	12,18%	1,88%	0,91%	0,52%	0,28%	0,01%	0,01%	0,01%

Ábra 5. A laikus közvélemény korpusz forrásainak darabszáma és százalékos megoszlása. Forrás: saját szerkesztés

[2] A korpuszokhoz használt keresés: ((kivándor*) AND NOT (ukrán*) AND NOT (Ukrajn*) AND NOT (zsidó*) AND NOT (Izrael*))

A másik korpusz a hazai szerkesztett média, tehát a hazai sajtóorgánumok honlapjain közölt írásokból tevődik össze. Ez a legyűjtés után 1 425 elemből, majd az adattisztítás után 1 148 (megegyező azonosító kiszűrése), illetve 1 085 elemből (ismétlődő elemek kiszűrése) állt. A következő részben e korpuszok feldolgozásának és elemzésének lépései kerülnek bemutatásra.

4.2. Eszközök és eljárások

Az elemzést kevert módszertannal végeztemem. A módszertanhoz két automatizált szövegelemzési eszközt használtam. A MEH szöveg (.txt) vagy más tagolt táblázatos formában (például .csv) tárolt szövegeket tud beolvasni. E dolgozat korpuszát szöveg fájlok alkotják. Minden egyes legyűjtött hozzászólás, bejegyzés, cikk egy szövegfájlban szerepel, azonosítóval ellátva. Ahogy az fentebb kifejtésre került, az MEH eszköz különböző menüpontjaiban a korpusz előfeldolgozásának lépései állíthatóak be. Az átváltási és a tiltószó listánál az eszköz magyar nyelvű listái kerültek betöltésre, tartalmuk szerkesztése nélkül. E listák alapján történt a lemmatizálás, illetve a tartalommal nem bíró szavak kiszűrése. A szövegek kisbetűsített formában kerültek elemzésre. A tíznél kevesebb szó alkotta szövegek nem képezték ennek részét. A rövidebb szövegek elsősorban az eredmények jobb kezelhetősége miatt kerültek kizárásra. Ezt tehát az előző oldalon leírtak mellett egy másik módja volt a nagy mennyiségű szöveges adat szűrésének. A hosszabb szövegek továbbá alkalmasabbak a vélemény kifejtésére is. Az elemzés egységei az egyes szavak, tehát az 1-gramok, közülük pedig csak azok, amelyek legalább ötször szerepelnek az adott korpuszban.

A beállítások után az MEH két olyan kimeneti fájlt generált, amelyek további elemzésre alkalmasak. Az egyik a dokumentum-kifejezés mátrix, amely a korpusz témáinak kinyeréséhez szükséges. A mátrix SPSS-ben, szöveges fájlként került megnyitásra. Az MEH eszköz bemutatásánál szereplő tanulmányokkal megegyezően, e dolgozatban is dimenziócsökkentést végzek a dokumentum-kifejezés mátrixon. A dimenziócsökkentés az MEM eljárásnak megfelelően PCA-vel történik, ahhoz pedig, hogy az egyes faktorok ne álljanak korrelációban egymással, varimax rotációt alkalmazok. Az MEM eljárást alkalmazó más tanulmányok mintájára (Markowitz, 2021; Pham és mtsai, 2023), a főkomponensek sajátértékei alapján, a laikus közvélemény korpuszból három, az online sajtó korpuszból nyolc főkomponenst tartok meg.

Az egyes főkomponenseket a laikus közvélemény korpusz esetében azon szavak alkotják, melyek faktorsúlya legalább 0,30. Ez a határérték a nagy elemszámú, rövid szövegeket tartalmazó korpuszok esetében ajánlott (Markowitz, 2021). Az online sajtó korpusz esetében -amely kisebb elemszámú és hosszabb szövegekből áll- a főkomponenseket a minimum 0,20 faktorsúlyú szavak alkotják. Ezek a határértékek tehát, a MEM eljárás szerint, a korpuszok jellemzőihez vannak igazítva (Chung és Pennebaker, 2008; Boyd, 2017).

A főkomponensek számát a sajátértékek alapján határoztam meg, a hüvelykujjszabályt figyelembe véve. Eszerint, a legalább 1 sajátértékű főkomponenseket tartottam meg (Landau és Everitt, 2004). A sajtó korpusznál a főkomponensek az összes megmagyarázott varinaciahányad 11 százalékát teszik ki. A laikus közvélemény korpusz esetében a főkomponensek az összes megmagyarázott varinaciahányad 5 százalékát teszik ki. A megőrzött információmennyiség a klasszikus kutatásoknál túl kevés lenne mindkét korpusz esetében. A szövegelemzésnél azonban ez megszokott, ugyanis több látens téma is megfigyelhető. A laikus közvélemény korpusz esetében a mindössze 5 százalékos megőrzött információmennyiség arra utal, hogy nincsenek nagy, meghatározó témák. Mindezt követően, a főkomponensek, az őket alkotó szavak alapján felcímkézésre kerültek, az MEM eljárást követve. Ezek a címkék felelnek meg a korpuszokban megjelenő fő témáknak.

Az MEH eszköz másik kimeneti fájljában, a Gyakorisági Listán, a szavak gyakoriság alapján növekvő sorban szerepeltek. Annak első 50 szava közül azok alkották az elemzés tárgyát, melyek az elemzés szempontjából jelentéssel bírnak (9. ábra). E szavakat az AntConc eszköz Kollokáció, N-Gram és Klaszter funkcióival tovább elemeztem. Az AntConc funkcióihoz így az MEH által előállított Gyakorisági Lista első 50, jelentéssel bíró szavát adtam meg. Mivel az AntConc eszköz nem alkalmaz előfeldolgozást az elemzés előtt, keresőszóként a szavak szótövét és egy „*” karaktert adtam meg. Ez a helyettesítő karakter bármennyi karaktert helyettesíthet, így a keresési találatok között szerepel az összes olyan szó, ami a megadott szótővel kezdődik. Az ékezetes betűk helyett egy másik helyettesítő karaktert, a „?” karaktert is használtam, ami pontosan egy karaktert helyettesít. Így például, a „külföld”, a „magyarország” vagy az “itthon” szavak összes, ragozott, esetlegesen ékezet nélkül írt változatát a „k?lf?ld*”, a „magyarorsz?g*” és az „itthon*” kifejezéssel kerestem.

Az eredményeket a Konkordancia, illetve a Fájl nézet funkciókkal ellenőriztem a kontextus alapján és ugyanezen funkciókat használtam a példamondatok megadásához. Az

AntConc Keyness funkcióját használva összevetettem a laikus közvélemény és az online sajtó korpuszt. A két korpuszt mind referencia, mind célkorpuszként használtam, egymással felcserélve. Ezzel megkaptam azon szavakat, amelyek az adott korpuszban szignifikánsan gyakrabban fordulnak elő a másik korpusz szavaihoz képest. Az alapbeállításokon nem változtattam. Az így kapott listák első 100 szavát elemeztem.

Első részében a PCA által kinyert témák és az AntConc Keyness funkciójának eredményei alapján mutatom be a laikus közvélemény és az online sajtó kivándorlás körüli diskurzusainak fő témáit, kiegészítve azok nyelvi sajátosságaival. Az elemzés második részében, a laikus közvéleményre koncentrálva, az ott megjelenő diskurzusokat, azok jellemzőit mutatom be. Ezt az MEH Gyakorisági Listáján elöl szereplő szavak AntConc-kal való elemzése alapján teszem. Az elemzés ismertetése után a használt automatizált szövegelemzési eszközöket értékelem.

5. Eredmények

5.1. Kivándorlás-diskurzusok

5.1.1 A laikus közvélemény és az online sajtó jellemzői és témái

A laikus közvéleményre jellemző az olyan szavak használata, mely a személyes vélemények, tapasztalatok, érzések megosztására utal. Az „én", „nekem", „szerintem", „értem", „remélem" szavak azt mutatják, hogy az egyéni nézőpontok előtérbe kerülnek a kivándorlásról szóló diskurzusokban. A laikus közvélemény tehát saját nézőpontjából ír a kivándorlásról, annak okairól, következményeiről. Az egyén személyes élményei és értelmezési keretei így hatással vannak e jelenséggel kapcsolatban megosztott véleményére. Ezzel szemben, a sajtóban megjelenő „szerint” és „körében” szavak nem egy személyes megközelítésre utalnak. A kivándorlással kapcsolatban közölt információk, elemzések vagy állítások mögött valamilyen forrás, más emberek, más intézmények véleménye vagy egy kutatási eredmény áll. A cikkek szerzője tehát csak közvetítő szerepet tölt be és nem saját véleményét osztja meg. A sajtóban megjelenő diskurzusok így inkább mutatnak egy szakmai képet. Ezt egészítik ki a „százalék”, „aránya”, „százaléka” „száma” szavak. Az online sajtó diskurzusban fontos szerepe van az empirikus adatoknak. A kivándorlást tárgyaló cikkekben a

hangsúly a személyes vélemények helyett a tényeken alapuló közlésen, magyarázaton van. E cikkek tehát elemzik vagy adatokra támaszkodva mutatják be a kivándorlást.

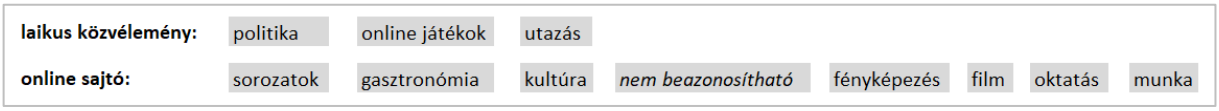
A sajtó tárgyilagosságához és a közösségi média informalitásához köthető megfigyelés, hogy a laikus közvélemény gyakrabban használ érzelmeket, elégedetlenséget kifejező szavakat. A „buta" és „szar" szavak gyakori megjelenése arra utal, hogy a kivándorlással kapcsolatos diskurzusokban erős negatív érzelmek, elégedetlenség és frusztráció is megjelenik. Az ilyen szavak megjelenése megerősítheti azt a korábbi megállapítást, amely szerint a vizsgált időszak egy érzelmileg túlfűtött periódus volt. Ez tehát a kivándorlással kapcsolatos diskurzusokban is megfigyelhető. A laikus közvélemény a platform sajátosságai miatt reagálhat az egymás által leírtakra, így a kivándorlás okairól, következményeiről vagy megítéléséről különböző véleményen lévő emberek között kialakuló viták is eredményezhetnek ilyen szavakat.

Emellett a közösségi médiában szintén gyakrabban megjelenő szavak az „ön", „te", „neked", „tudod". Ezek az emberek közötti interakcióra, a másik megszólítására utalnak. Míg az online sajtó tárgyilagosan közli az információkat, tájékoztat, egy általános közönséghez szól, addig a közösségi médiában az emberek reagálhatnak egymás gondolataira. A kivándorlással kapcsolatos diskurzusok során, így kialakulhatnak érvrendszerek, az embereknek célja lehet, hogy meggyőzze a másikat igazáról. A laikus közvélemény írásaira továbbá jellemzőbb a töltelékszavak használata (6. ábra). Ezen írások hangneme informális, nem annyira kötött, mint az az online sajtóban tapasztalható. Ez az informális kifejezésmód jellemzőbb a hétköznapi beszédre is. Ez utóbbi a leírt gondolatok szerkesztésének hiányából is adódhat.

Ábra 6. A két korpusz szóhasználatának legjelentősebb különbségei. Bal oldalt az online sajtókorpuszban, jobb oldalt a laikus közvélemény korpuszban szignifikánsan gyakrabban előforduló szavak szófelhője látható. Forrás: Az AntConc Wordcloud funkciójával készített szófelhők az AntConc Keyness funkciójának eredményei alapján

A vizsgált időszakban az online sajtóban megjelenő kivándorlással kapcsolatos írások egyik fő jellemzője, hogy ezen cikkekben a diskurzus nem a kivándorlás témájában folyik. Nem annak különböző, például politikai, gazdasági, társadalmi vagy akár kulturális aspektusairól írnak. A „című", „film", „fotó", „rendező", „díjas", „fesztivál", „közönség", „színház", „premier", „oscar", „klasszikus" szavak, melyek gyakrabban fordulnak elő az online sajtó diskurzusiban, mind a kultúra témaköréhez kapcsolhatóak. Ezekben a cikkekben, bár megjelennek a kivándorlás kulturális hatásai, például filmek, színházi előadások vagy művészeti események kapcsán, maga a kivándorlás azonban háttérbe szorul. Az ilyen cikkek elsősorban művészek életrajzai, vagy program-, illetve filmismertetők, ahol kitérnek az egyes alkotók életútjára. Ezen életutak egyik eleme a kivándorlás. Az ilyen cikkekben tehát nem a kivándorláson van a hangsúly, hanem magán az az életúton, valamint az egyes személyeken. A közösségi médiában ezzel ellentétben gyakrabban fordulnak elő olyan szavak, melyek a fogadó országra („kint"), a kibocsájtó országra („itthon", „itt", „haza"), a kettő közötti mozgásra („megy", „jönnek", „mennek", „menjen", „menjenek") illetve magára a kivándorlásra utalnak („kivándorol", „kivándorolni"). A kivándorlás-diskurzusok tehát a sajtóban kevéssé jelennek meg, háttérbe szorulnak, míg a laikus közvélemény aktívan foglalkozik a kivándorlás témájával a 2022-es országgyűlési választásokat követő hónapokban.

A laikus közvélemény kivándorlás-diskurzusaiban továbbá megjelennek a magyar politikához, illetve a gazdasághoz köthető szavak („fidesz'", „gyurcsány", „orbán", „eu", „dolgozni"). A kivándorlás így összefonódik politikai és gazdasági döntésekkel, kérdésekkel. Mindez a vizsgált időszak tekintetében azt jelenti, hogy a jelen lévő társadalmi, politikai és gazdasági feszültségek kiéleződése a kivándorlás-diskurzusokban is megnyilvánult. Ez azonban az online sajtó kivándorlással kapcsolatban megjelenő cikkeire kevéssé jellemző.



Ábra 7. A két korpuszban megjelenő nagyobb témák.
Forrás: saját szerkesztés a főkomponenseket alkotó szavak alapján

Az online sajtó-diskurzusok fő témái, ahogyan az a 7. Ábrán látható, a kultúra, fényképezés, a filmek, a sorozatok és a gasztronómia. Emellett a munka és az oktatás is megjelenik. Ez utóbbi elsősorban a felsőoktatás, az informatikai képzések és a kivándorló

szakemberekkel kapcsolatban. Ezzel szemben, a laikus közvélemény kivándorlás-diskurzusai a politika és az utazás témák köré csoportosulnak. Továbbá, megjelenik egy másik nagyobb téma is, az online játékok, amit a korpuszba magas számmal bekerülő hirdetések okozhatnak. A következőkben a laikus közvélemény kivándorlás-diskurzusai kerülnek részletesebb bemutatásra, ahol a nagyobb témák megjelenése kevéssé volt jellemző.

5.1.2. A laikus közvélemény kivándorlás-diskurzusainak bemutatása

A laikus közvélemény kivándorlással kapcsolatos diskurzusainak több központi elme is van a 2022-es magyarországi országgyűlési választásokat követően. E központi elemek az egyéni értékek, a politika és a gazdaság. Ezek bemutatása következik a leggyakrabban használt szavak és azok kontextusa alapján.

„Sokkal többre vittünk mint azt valaha képzeltük."

Nagy véleménykülönbség mutatkozik azzal kapcsolatban, hogy az ország vagy az egyéni anyagi jólét a fontosabb. A laikus közvélemény szerint e kettő, ha nem is teljesen, de konfliktusban áll. E két érték nagyobb skálára kivetítve a társadalmi, közösségi, nemzeti érdekek és az önös érdekek ütközésében jelenik meg. A diskurzusokban megjelenő vélemények alapján a kivándorló magyarok a saját boldogulásukat előre helyezve döntenek úgy, hogy elhagyják az országot. Tapasztalatik vagy reményük szerint külföldön jobb karrierlehetőségek, magasabb fizetések és nagyobb megbecsültség a jellemző. A magyarországi élethez képest a kivándorlás egy olyan lépés, amely számos lehetőséget teremt. A gazdasági előrejutás érdekében a haza iránti hűség nem prioritás. Az ország elhagyása az anyagi boldogulás reményében elsősorban hosszútávú tervek része. Az, hogy a gyerekek más körülmények között nőhetnek fel, a diskurzusban résztevevő egyes családosok számára döntő fontosságú. Véleményük szerint külföldön a jobb oktatási rendszer megléte és egy egészségesebb, nem gyűlölködő környezet, jelentős pozitív hatással van vagy lehet a gyerekek jövőjére.

„A magyarságot előtérbe helyező értelmiség nélkül csak zuhanunk lefelé."

Mindezek ellentéte is megjelenik a laikus közvélemény diskurzusaiban. Az anyagi jólét választása és emiatt a haza elhagyása egyesek szerint a kapzsiságot jelenti. A külföldi élet gazdasági lehetősége nem anyagi biztonságot, hanem többletjövedelmet teremt.

Magyarország, a haza és a nemzet sorsa számukra nem elsődleges. A kivándorló embereknek felelőssége van az ország jelenlegi helyzetében, ugyanis nem vesznek részt annak jobbá tételében, a problémáktól inkább elszigetelik magukat. Az ország helyzete tehát nem olyan fontos számukra, nem érdekeltek annak jobbá tételében. A kivándorlók felismerik és tapasztalják a problémákat és változást szeretnének, éppen emiatt hagyják el az országot. Azonban, az ilyen emberek kivándorlása az ország helyzetét érintő, pozitív irányba történő változásokat lassítják vagy ellehetetlenítik.

A diskurzusokban a kivándorlás mellett döntő emberek felelőssége elsősorban a nyugdíj és annak előállításával kapcsolatban jelenik meg. A külföldön munkát végző magyarok nem termelnek adóbevételt az országnak. A jelenlegi, valamint a jövőbeli nyugdíjasok helyzetére így kihat a külföldre költözésük. A kivándorlás mellett döntő emberek tehát az ország más polgáraira nincsenek tekintettel. Saját boldogulásuk előtérbe helyezése, ugyan közvetettem, de befolyásolja a jelenlegi és jövőbeli magyar idősek megélhetését.

A kivándorlók közül továbbá kiemelt felelőssége van a képzett munkaerőnek és az értelmiségi rétegnek. Az ő kivándorlásuk egy hatalmas vesztesége az országnak. Bennük ugyanis meg lett volna, vagy meg lenne a potenciál a pozitív változások eléréséhez. Közülük vagy gyerekeik közül kinőhetne egy új vezetői réteg és hozzáértő szakemberek is nevelkedhetnének. Számukra azonban Magyarország és az ott élő emberek jóléte nem elsődleges, saját boldogulásukat helyezik ezek elé. Az „igaz szívű", „öntudatos" emberek kitartanak az ország mellett, bármi is történjék. Az olyan magyarok, akikben „van tartás", nem hagynák el hazájukat. Az önös érdeket előtérbe helyező, elvándorlás mellett döntő magyarok így bizonyságot tesznek arról, hogy nem „igaz magyarok". A diskurzusokban ennek megítélése többféle. Az ország számára ez jelenthet veszteséget, például a szakemberek hiánya miatt. Mindemellett azonban az a vélekedés is megjelenik, amely szerint az országban nem kívánatosak az olyan emberek, akik nem tartanak ki hazájuk mellett.

„Ott sem kolbászból van a kerítés"

A laikus közvéleményben egy másik nagyon gyakori elem a tévhitek megjegyzése a külföldi élettel kapcsolatban. Megjelenik az a vélekedés, hogy azok, akik szeretnének kivándorolni Magyarországról, túl sokat várnak a költözéstől. Külföldön sem működik minden jól, az új élet megteremtésével nagyon sok új probléma is jár. Ez utóbbi megállapítást a már

kint élő magyarok is megerősítik. Maga a kivándorlás önmagában nem oldja meg az ember problémáit. Akik külföldön érvényesülni tudnak, itthon is tudnának. Ellenben, akiknek itthon nem sikerült, azokat előreláthatólag a kinti életükkel kapcsolatban is csalódások fogják érni. Olyan elvárásokkal mennek ki, melyek nem felelnek meg a valóságnak, tévhiteik vannak a külföldi élettel kapcsolatban.

„szerencsére az ország előre megy nem hátra!"

A vizsgált szövegek egyik visszatérő, fő témája a politika. Az ezzel kapcsolatos diskurzusokban a kivándorlás nem minden esetben, mint központi elem jelenik meg. Azonban, a kivándorlás összefonódik a politikai diskurzusokkal. Az ilyen diskurzusok visszatérő elemei a 2022-es országgyűlési választások kampányszövegei, valamint azok átírása. A „Magyarország előre megy, nem hátra", illetve a „Magyarország jobban teljesít" szlogenek mind az adott időszak jellemzői. Az olyan diskurzusok, melyekben ezek megjelennek, politikai témájúak, kormánykritikusak. E szlogenek ironikus használata a kormány rossz döntéseire és annak következményeire utal. Ezen vélekedések szerint, bár a kormány azzal kampányol, hogy az ország jobban teljesít, a saját tapasztalatok ezt nem támasztják alá. A kivándorlás is az elhibázott döntések egyik tünete. Az ország elhagyása egy radikális döntés, amely meghozatala lemondásokkal jár. A kivándorlás tehát az ország helyzetével való elégedettlenség egy visszajelzési formája is.

„Aki itthon marad az az alja, jobbágy"

A kivándorlás tehát a jelenlegi kilátástalanságra, az országában való csalódásra is megoldásként jelenik meg a politikai és gazdasági támájú diskurzusokban. Egyes vélekedések szerint az országban gyökeres változás szükséges. A jelenlegi helyzet és a politikai döntések nem adnak okot egy jobb jövő reményére. Az állami pénzek rossz felhasználására, korrupcióra hivatkozva úgy érzik, a magyarországi adó befizetése nem járul hozzá az ország előre lepéséhez. Az oktatás és az egészségügy állapota nem megfelelő. Az adójuk befizetésével nem szeretnének hozzájárulni a pénzek ilyen felhasználásához. Továbbá, nem is szeretnének egy olyan országban élni, ami rosszul működik. Ők jobb életkörülményeket, több lehetőséget szeretnének. A kivándorlás ebben a megvilágításban egy jó alternatíva lehet arra, hogy az ember egy jobb környezetben folytassa életét és teljesebb életet éljen. Így azok, akik emiatt el tudják hagyni Magyarországot, úgy érzik, jó helyzetmegítélő képességük van. Képesek arra,

hogy belássák a problémákat és elég bátrak ahhoz, hogy lépéseket is tegyenek egy másfajta élet felé. Ezzel szemben, akik az országban maradnak, nem veszik észre saját helyzetüket és nem ismerik fel azt, hogy lehetne jobb életük is. Az ilyen emberek a kivándorlók szerint képzetlenek, alacsony iskolai végzettséggel rendelkeznek és emiatt könnyen manipulálhatóak. Az ő felelősségük is kiemelt abban, hogy nem lehet változást elérni.

„Akinek nem tetszik, el lehet menni"

A Magyarországgal szembeni eltúlzott kritika bírálata is megjelenik a diskurzusokban. Az ország és annak kormánya, bár nem tökéletes, mégis működőképes. Az ország történelméből vett példák, egyes vélekedések szerint azt támasztják alá, hogy korábban sem volt jobb az ország helyzete. Az országgal szembeni erős kritikák nem vezetnek előre. Az otthoni helyzetet nem bírálni kell, hanem megbecsülni. Azok, ugyanis, akik arra törekvéseket tesznek, bármilyen helyzetben boldogulnak. Azok számára, akik nem tudják megtalálni a pozitívumot jelenlegi helyzetükben, és nem tudnak változtatni sem ezen, a kivándorlás egy megoldásként szolgálhat. A diskurzusok e központi elemei mellett megfigyelhetőek a szövegek különböző jellemzői is, melyek a témák bemutatásánál leírtakat egészítik ki.

„ön buta. Nagyon buta."

A 2022-es országgyűlési választásokat követő időszakban a laikus közvélemény kivándorlással kapcsolatos diskurzusira jellemző az indulatos szavak használata és a személyeskedés. Ezek elsősorban a választások miatti politikai feszültségekből adódnak. A nem egyező vélemények bizonyos politikai, gazdasági döntésekről, illetve a kivándorlás megítéléséről, vitákhoz vezet.

„Nagyon sok ismerősöm él külföldön"

A diskurzusok során a kivándorlásról alkotott szubjektív vélemények kifejtésekor megemlítésre kerülhet a személyes érintettség. Ezáltal a leírtak további hitelességet nyernek, hiszen forrásuk egy olyan ember, aki személyes kapcsolatai vagy tapasztalatai miatt többlet tudással rendelkezik az adott jelenségről. Referenciakén szolgálhatnak például az ismerősök, családtagok által elmondottak, de a saját külföldi vagy magyarországi élettel kapcsolatos tapasztalatok is megosztásra kerülnek.

5.2. A no-code eszközök limitációi

Az előző részben tárgyaltak mellett a no-code megoldás hátrányait és az eszközök limitációit is fontos kiemelni, ugyanis azok nagyban meghatározták az eredményeket. Egyik vizsgált automatizált szövegelemzési eszköznél sem volt lehetőség arra, hogy a korpusz ismétlődő elemei eltávolításra kerüljenek. Bár mindez megoldható még az eszközök használata előtt, e folyamat időigényes. Miután ez megtörtént, tehát kiszűrésre kerültek a teljesen ismétlődő szövegek, a korpuszokban továbbra is sok, közel azonos tartalmú szöveg maradt. Ez elsősorban az online sajtó korpuszra volt jellemző. Az AntConc eszköz Konkordancia funkciójának segítségével bebizonyosodott, hogy többi funkciónál - például az N-Gram funkciónál - az együtt szereplő szavak egy jelentős része azért került a Gyakorisági Lista elejére, mert a korpuszban közel azonos tartalmú szövegek szerepelnek. Témavezetőm tapasztalatai szerint, ez más eszközök számára is probléma online sajtó korpuszok esetén. A számos átvett vagy némileg átfogalmazott szöveg, például topikelemzésnél akár, meghatározhatják az egyes topikokat. Jelen dolgozatban az átvett tartalmak azért jelentenek problémát, mert az N-Gram funkció azt vizsgálja, melyek azok a kifejezések, amik több kontextusban is együtt szerepelnek. Az eredmények azonban, az ismétlődő szövegek miatt nem voltak hasznosak a vizsgálat e szempontjából.

Az a kísérlet tehát, hogy a megegyező tartalmak ne szerepeljenek az elemzésben, nem tudott teljesülni. A korpuszban több szöveg tartalma is szóról szóra megegyezett. Közöttük olyan apróbb változások voltak, mint például egy szöveg végén lévő forrás link, képek forrásának megnevezése vagy a tartalmat jelölő címkék felsorolása. Egy, az ismétlő tartalmak torzítása nélküli eredményhez, a korpusz szövegeinek további szűrésére lenne szükség, amely már általános felhasználói tudással nem teljesíthető. Ilyen szűrések programozással egyszerűen kivitelezhetőek. Például, ha csak azon szövegeket tartjuk meg, melyekben az egymás után közvetlenül előforduló szavak legalább egy megadott, magas (akár 90 vagy 80) százalékban eltérnek egymástól. Egy másik módszer egy hasonlósági mutató (Jaccard similarity) definiálása lenne, amely az egyes szöveg páronkénti összevetéséhez alkalmazható. Amely szövegeknél e mutató egy küszöbél nagyobb hasonlóságot mutat, ott az egyik random választással kikerül a korpuszból.

A korpusz előfeldolgozása során keletkezett hibákat folytatva, az MEH eszköz limitációi kerülnek ismertetésre. Az MEH eszköz egyik nagy előnye, hogy a Stopszó Lista és az Átváltási

Lista magyar nyelvű változata is megtalálható a programban. E beépített listák azonban, ahogyan az a Gyakorisági Listán szereplő szavakon is látszik, nem teljes körűek. A listák e hiányosságai megnehezítik az eredmények értelmezését, ugyanis sok olyan szó hiányzik belőlük, amely tartalommal nem rendelkezik, ez által pedig irreleváns az elemzés szempontjából. Mindemellett maga az elemzés végeredménye is torzulhat az által, hogy az azonos jelentésű szavak különböző formában többször, és nem összeadódva kerülnek a listára.

Az MEH eszköznél az Átváltási Lista a lemmatizálás alapja. A lemmatizálás tehát nem nyelvtani szabályok, például a tárgyragok levágásával, hanem egy lista alapján történik. E listán szereplő szavak ideális esetben a korpusz összes szavát át tudnák váltani annak szótári alakjára. Így a feldolgozott korpuszban már nem szerepelnének olyan szavak, melyeknek toldaléka van. Az MEH eszköz által létrehozott Gyakorisági Listát alkotó szavakra tekintve azonban észrevehető, hogy ez nem teljesül. Az alábbi, 8. ábrán látható egy példa arra, hogy a „magyarország” szótővel rendelkező szavak hány különböző elemmel, milyen gyakorisággal kerültek az online sajtó korpuszban előforduló szavak Gyakorisági Listájára. A szövegelemzés kódolást igénylő megoldásai között a nyelvtani szabály alapú lemmatizálás megbízhatóbb eredmények hoz. A HuSpaCy (Orosz és mtsai, 2022) például egy olyan eszközkészlet, amely a magyar nyelvű szövegek előfeldolgozását képes hatékonyan elvégezni. A lemmatizálás során az egyes szavakhoz szófajcímkéket rendel és ez alapján ismeri fel, majd távolítja el a szavak végződését, kinyerve ezzel a szavak szótári alakját.

Szó	Gyakoriság
magyarország	956
magyarországot	62
magyarországért	28
magyarországhoz	14
magyarországból	9
magyarországban	6
magyarországtól	5
magyarországa	4
magyarországán	4
magyarországát	3
magyarországé	3
magyarországi	2
magyarországig	2
magyarországinak	2
magyarországés	2
	1102

Ábra 8. Az online sajtó korpusz Gyakorisági Listáján szereplő „magyarország" kezdetű szavak.
Forrás: saját szerkesztés az MEH Gyakorisági Listája alapján

Mindemellett, a 8. ábrán megfigyelhető az Átváltási Lista egy másik hiányossága is. A Gyakorisági Listán bizonyos szavak helyesen, ékezettel írt alakja, valamint ékezetek nélkül írt alakja is külön szerepel, ami szintén egy probléma. Az ékezetpótlás, ami a magyar nyelvű, elsősorban közösségi médiában megjelenő szövegek esetében fontos, az MEH eszköztárában nem található meg. Erre magyarázat lehet, hogy az eszközt elsőként angol nyelvű szövegek elemzésére fejlesztették ki, amelyben nem használnak ékezeteket. Az attól eltérő nyelvű szövegeket az eszköz tehát nem tudja pontosan előkészíteni. Az ékezetpótlás kódolással könnyen javítható lenne, az MEH eszköznél azonban csak az Átváltási Lista manuális kiegészítésével lehetséges. Ez utóbbi egy nagyon időigényes folyamat, ugyanis a korpuszban szereplő összes, a helyesírási szabályoknak nem megfelelően, ékezet nélkül írt szó mind az Átváltási Listára kellene, hogy kerüljön, kiegészítve a helyesen írt változatával. Mindez lerövidíthető azzal, ha csak a Gyakorisági Lista elején - például az első 50 vagy 100 szó között - helyet kapó szavak esetében kerülnek manuálisan javításra az ékezet nélkül írt szavak. Ezek ugyanis az elemzés szempontjából a fontosabb elemek. Mindent összevetve azonban ez továbbra is egy olyan időigényes folyamat, ami nem zárja ki a torzítás lehetőségét.

Az olyan helyesírási hibák és elgépelések, mint például az ékezetek elhagyása, a közösségi médián megjelenő szövegek egyik jellegzetessége. E dolgozat laikus közvélemény korpuszban a helyesírási hibák és elgépelések száma alacsony. A leggyakrabban előforduló elgépelt szó az „os”, amely 55 alkalommal fordul elő, ezzel a 236. leggyakoribb szó. Bár a dolgozat eredménye nem torzul jelentősen a helyesírási hibák és elgépelések miatt, mégis érdemes megemlíteni, hogy kódolással megoldható a magyar nyelvű szövegek helyesírási hibáinak automatikus javítása (Siklósi és mtsai, 2016).

A Stopszó Lista az MEH eszköz egy másik olyan listája, amely, bár elérhető magyar nyelven, mégis nagyon hiányos. E listára azon szavak kerülnek, melyek nem rendelkeznek jelentéssel, emiatt pedig kiszűrésre kerülnek az elemzésből. A 9. ábrán látható, hogy a laikus közvélemény korpusz Gyakorisági Listájának első 50 szava között is előfordulnak a kutatáshoz nem releváns szavak, például kötőszavak. Emellett, sok személynév is bekerült az elemzésbe a közössági oldalakon való megjelölések miatt. A Stopszó Lista hiányossága az Átváltási Lista hiányosságához hasonló problémákat jelent. Az MEH eszköz által létrehozott eredmények értelmezése több időt és több koncentrációt igényel a nem releváns tartalmak manuális

kiszűrése miatt. Mivel a Stopszó Lista szabadon szerkeszthető, így annak tartalmát lehetséges a kutatáshoz illeszkedve kiegészíteni módosítani.

Sorszám	Szó	Gyakoriság	Sorszám	Szó	Gyakoriság
1	**&#65533;**	**1080**	26	vesz	243
2	**tud**	**981**	27	**es**	**240**
3	magyar	866	28	**–**	**235**
4	ember	761	29	kivándorlás	234
5	megy	756	30	idő	229
6	ország	712	31	élet	222
7	**fog**	**540**	32	gyerek	218
8	nagy	525	33	**áll**	**214**
9	él	486	34	haza	214
10	**akar**	**471**	35	**ír**	**209**
11	dolgoz	413	36	saját	209
12	mond	366	37	**szó**	**208**
13	jön	360	38	vissza	205
14	**tesz**	**333**	39	fiatal	204
15	marad	329	40	fizet	201
16	magyarország	321	41	kormány	198
17	lát	321	42	**kap**	**195**
18	itthon	292	43	fidesz	194
19	munka	277	44	rész	194
20	pénz	263	45	kint	191
21	külföld	261	46	…	189
22	gondol	260	47	biztos	182
23	igaz	249	48	kis	181
24	**ad**	**247**	49	online	181
25	**annyi**	**245**	50	**s**	**177**

Ábra 9. A laikus közvélemény korpusz Gyakorisági Listájának első 50 eleme. A kutatás számára nem releváns szavak vastaggal jelölve. Forrás: saját szerkesztés az MEH eredményei alapján.

E feljebb tárgyaltak hiányosságokra a 9. ábrán egy példa is látható. Bár az „és” szó szerepel a Stopszó Listán, két másik alakja mégis bekerült az elemzésbe. Az „es” szó az „és” ékezetlenül írt változata. Az a hiba, hogy e két szó mégis külön szerepel, ékezetpótlással javítható lenne. Az „s” szintén az „és” szó egy másik változata, amely az Átváltási Lista kiegészítésével javítható lenne. Amennyiben mindez megtörténik, mind az „es”, mind az „s” szavak a Stopszó Listára „és” formába kerülve, kimaradtak volna az elemzésből.

E dolgozatban az MEH eszköz használatával az MEM eljárást követtem. Az eljárás utolsó lépéséhez szükséges PCA eredménye nem az összes főkomponens esetében volt értelmezhető. Az egyik főkomponens olyan szavakat tartalmazott, amelyekből nem olvasható ki egy egységes téma. E főkomponenst azon szövegek elolvasása után lehet értelmezni, melyekben a főkomponenst alkotó szavak előfordulnak. A felcímkézéshez így egy mélyebb, kvalitatív elemzés szükséges. Ez azonban nem része az MEM eljárásnak, így e dolgozatnak sem.

Míg az MEH eszköznél a korpusz előfeldolgozásának hiányosságairól lehetett írni, az AntConc eszköznél az előfeldolgozás egyáltalán nem lehetséges. A lemmatizálás és stopszavazás tökéletlenségéből eredő problémák fentebb már kifejtésre kerültek. Mindez az AntConc összes funkciója használatakor megnehezíti az elemzést. E dolgozathoz a lemmatizálásból adódó hibákat a „*” helyettesítő karakter használatával próbáltam kiküszöbölni, így azonban bizonyos esetekben nem releváns találatok is kijöttek, melyek az adott keresőszóval kezdődtek.

Az AntConc eszköz azonban mindezek mellett jól használható, megbízható. Az MEH eszköz ellenben a dolgozat során végzett teszteken több alkalommal is pontatlan elemzést végzett. Ezek a problémák elsősorban az n-gram beállításokkal voltak kapcsolatosak. Bár a szavak gyakoriság alapú kiszűréséhez több választási lehetőség is tartozik, a próbák során ezek egyike sem működött megfelelően. Az egyik ilyen lehetőség, hogy az elemzésben csak a legalább x (megadható érték) alkalommal megjelenő szavak kerüljenek. E lehetőségnél a különböző értékek megadása nem befolyásolta a kimeneti fájlt. Tehát, hiába került beállításra az első próbaelemzésnél, hogy csak a legalább 1 000 alaklommal, a második próbaelemzésnél pedig csak a legalább 1 500 alkalommal megjelenő szavak képezzék az elemzés részét, a kimeneti fájlok ugyan azok lettek mindkét esetben.

A dolgozathoz is használt lehetőség, ami szerint a laikus közvélemény korpuszban legalább 5 alkalommal előforduló szavak kerüljenek az elemzésbe, szintén hibás eredményeket hozott. Az elemzés után létrejövő egyik kimeneti fájlban, a Gyakorisági Listán megfigyelhető, hogy az arra kerülő szavak egy jelentős része csak két, három, illetve négy alkalommal fordul elő a korpuszban. Szám szerint, a vizsgált szavak közül 7 371 szó szerepelt feleslegesen a listán, amely az összes listán szereplő szó közel 70 százalékát teszik ki. Ez a Gyakorisági Lista elemzését nem befolyásolta, ugyanis annak csak első 50 elemét vizsgálatm. Ezek a szavak azonban a dokumentum-kifejezés mátrixban is szerepeltek, ami a PCA során gondot okozott. Egyrészt, mivel az elemzésbe a szavak több, mint két harmada feleslegesen lett bevonva, maga a PCA jelentősen hosszabb lett. A mátrixokon végzett számítások erőforrás-igényesek, az említett probléma pedig még inkább megnöveli a számítási időt. Másrészt, olyan elhanyagolható szavak is bekerültek az elemzésbe, amelyek a korpuszt alkotó több, mint 7 000 szövegben mindössze 2 alkalommal fordult elő.

6. Konklúzió

6.1. Összegzés

Az információs társadalmat jellemző, folyamatosan termelődő adatokok új kapukat nyitottak meg a társadalomtudományok területén. Ezek ugyanis alkalmasak arra, hogy az emberek véleményéről, attitűdjéről képet adjanak, így korábban nem látott szélességét és mélységet kínálnak szociológiai elemzéseknek. Ugyanakkor, ez a hatalmas mennyiségű adat, amely rendelkezésre áll, újabb módszertani megoldásokat, elsősorban automatizálást igényel. Az online térben megtalálható szöveges információ feldolgozása komplex, több tudományterületen átívelő tudást igényel. Léteznek azonban olyan kódolást nem igénylő eszközök, melyek e módszert általánosan elérhetővé teszik.

E dolgozatban két olyan automatizált szövegelemzési eszközt mutattam be, melyek nem igényelnek programozói tudást. Ezen eszközök relevanciája az egyszerű használatukban rejlik. A Meaning Extraction Helper a szövegek előfeldolgozását végzi. Az eszköz több olyan beállítási lehetőséget is tartalmaz, amely lehetővé teszi, hogy a szövegek előfeldolgozása illeszkedjen az adott kutatáshoz. Előnye, hogy több más nyelv mellett, a magyar nyelvre is tartalmaz alapbeállításokat. Hátránya, hogy az alkalmazott megoldások egy része nem megfelelő, emiatt az eredmények többször pontatlanok. Mindemellett, az eszköz működése nem mindig megbízható.

Az AntConc elsősorban kulcsszavas keresés alapján képes szövegek elemzésére. Statisztikai számításokat alkalmazva elősegíti a szövegek mind kvantitatív, mind kvalitatív elemzését. Legnagyobb hátránya, hogy a szövegeket előfeldolgozás nélkül vizsgálja, így az eredmények nem szabhatóak az adott kutatáshoz. Előnye, hogy sok funkciója van, amelyek egymással interakcióban állnak, így megkönnyítve az eredmények értelmezését. További előnye a megbízhatóság.

Szakdolgozatomban e két eszköz használatával, kevert módszertannal végeztem kísérleti kutatást. Arra kerestem a választ, hogy milyen fő témák és diskurzusok jelentek meg a Magyarországról történő kivándorlással kapcsolatban a laikus közvéleményben és az online sajtóban a 2022-es év második felében. Ebben, az országgyűlési választásokat követő időszakban, a társadalmi és a politikai feszültségek kiéleződtek. A politikai élettel való elégedetlenség pedig a kivándorlás egyik fő oka, korábbi felmérések szerint. A kutatási kérdés

relevanciája, hogy a Magyarországról történő kivándorlás, az azt megelőző években csökkent, 2022-ben azonban jelentősen megnőtt. E jelenség vizsgálatához a laikus közvéleményben és az online sajtóban megjent írásokat elemeztem. Az online sajtó kevéssé foglalkozott a kivándorlás témájával a vizsgált időszakban. A laikus közvélemény kivándorlás-diskurzusainak fő, kutatáshoz kapcsolódó témái a politika és az utazás. E diskurzusoknak továbbá visszatérő eleme volt az ország, illetve az egyéni érdekek közötti konfliktus, a külföldi élettel kapcsolatos tévhitek, az országban való csalódás és az azzal szembeni eltúlzott kritika bírálata. Az online sajtó diskurzusaiban a kivándorlás háttérbe szorult, inkább kulturális, illetve gasztronómiai témák tárgyalása során kapott helyet. A kutatás számára releváns témák, melyek megjelentek, az oktatás és a munka.

6.2. További kutatási lehetőségek

E szakdolgozatban a kivándorlás-diskurzusok példaelemzése során alkalmazott megoldások nem mutatják be teljeskörűen a vizsgált eszközökben rejlő lehetőségeket. Lezárásként olyan más alkalmazási lehetőséget mutatok be, amelyek a téma mélyebb megértéséhez járulhatnak hozzá, illetve más korpuszok vizsgálatához mutathat példát.

A PCA alkalmazásával létrehozott főkomponensek elmentésével lehetőség van a kivándorlás-diskurzusok további vizsgálatára. Az adatbázis sorrendbe rendezése a főkomponensek alapján, lehetővé teszi azon szövegek lekérését, amelyek a legnagyobb abszolútértéket veszik fel. Így beazonosításra kerülnek azon tartalmak, amelyek a legnagyobb mértékben járulnak hozzá az adatok változékonyságához, legfontosabbak a kivándorlás-diskurzusokban megfigyelt nagyobb témáknál. Ez által, ezen szövegek kvalitatív elemzésére is lehetőség van. A dolgozatban alkalmazott szavak alapján való felcímkézéshez képest, ez egy gazdagabb, pontosabb elemzést tesz lehetővé.

Emellett, szintén a főkomponensek elmentése után, a két legnagyobb sajátértékű főkomponens síkján, szórásdiagrammal lehet ábrázolni az érintett szövegeket. Amennyiben az online sajtókorpusz egyes dokumentumai, forrásuk alapján felcímkézésre kerülnek, a diagramról leolvasható a témák előfordulása források szerint. Így tehát, vizuálisan is bemutatásra kerül, hogy a beazonosított nagyobb témákat, melyik hírportálok cikkei teszik ki. A felcímkézés nem csak az egyes hírportálok alapján történhet, hanem azokat csoportosítani is lehet, például politikai besorolásuk szerint.

Továbbá, a sajtó korpusz elemzésénél tapasztalt nagy számú, a téma szempontjából irreleváns tartalom is további szűrésre kerülhet a PCA által kapott főkomponensek alapján. Azon főkomponensek, melyek a kutatás szempontjából nem releváns témát jelölnek, szűrőként is használhatóak. E főkomponensekben nagy abszolútértékű faktorsúlyt felvevő elemek forrása beazonosítható. Ezen források pedig így kiszűrhetőek a korpuszból.

További kutatási lehetőség lenne, a szövegekben előforduló szavak alapján, a diskurzusokban előforduló kivándorlási célországok, illetve politikai szereplők beazonosítása. Ez mind az MEH Gyakorisági Listájával, mind az AntConc Szólista funkciójával kivitelezhető. Ebből kiindulva kvalitatív elemzéseket lehet végezni az érintett szövegeken, ugyanis mindkét eszköz használatával beazonosíthatóak azon szövegek, amelyekben a keresett szavak szerepelnek. Így tehát az eszközök a szövegek célzott szűrésére is alkalmazhatóak. Ezzel a módszerrel például a korábbi, célország-választással, illetve kivándorlást magyarázó tényezőinek vizsgálatával foglalkozó tanulmányokat (Blaskó és Fazekas, 2016) lehet kiegészíteni.

Mindezek mellett, a közélettel és politikával foglalkozó online hírportálok vagy politikai szereplők által közzétett közösségi média tartalmak alkotta korpuszok vizsgálatával be lehetne azonosítani a kormánypárti, illetve az ellenzéki retorikához kapcsolódó szavakat. Míg a laikus közvéleménynek teret adó tematikus blogok, közösségi média csoportok vagy videó feliratok automatizált elemzésével ezek köznyelvi lecsapódását lehetne vizsgálni.

Irodalomjegyzék

Anthony, L. (2005). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In: Anthony L. és mtsai. szerk., *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*. Tokyo, Waseda University. p. 7-13.

Anthony, L. (2023). Antconc (Version 4.2.4). [szoftver]. Elérhető: https://www.laurenceanthony.net/software/antconc.

Babbie, E. (2017). *A társadalomtudományi kutatás gyakorlata*. Budapest: Balassai Kiadó.

Barna, I. és Koltai, J. (2019). Attitude changes towards Immigrants in the turbulent years of the 'migrant crisis' and anti-immigrant campaign in Hungary. *Intersections. East European Journal of Society and Politics*, Vol 5, No 1, p. 48-70.

Bíró-Nagy, A. és Szabó, A. (2021). *Magyar Fiatalok 2021. Elégedetlenség, polarizáció, EU-pártiság.* Budapest: Friedrich-Ebert-Stiftung.

Bíró-Nagy, A. és mtsai (2022). *Széttartó világok: Polarizáció a magyar társadalomban a 2022-es választások után.* Budapest: Friedrich-Ebert-Stiftung-Policy Solutions.

Zsuzsa, B. és Gödri, I. (2014). Kivándorlás Magyarországról: szelekció és célország-választás az „új migránsok" körében. *Demográfia*, Vol 57, No 4, p. 271-307.

Boyd, R. L. (2017). Psychological text analysis in the digital humanities. In: Hai-Jew, S. szerk., *Data analytics in the digital humanities*. New York: Springer International Publishing. p. 161–189.

Boyd, R. L. (2022). MEH: Meaning Extraction Helper (Version 2.3.00). [szoftver]. Elérhető: https://www.ryanboyd.io/software/meh.

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*. Vol 29, No 3, p. 509-553.

Chomsky, N. (1957). Syntactic structures. Berlin: Mouton.

Cohn, M. A., és mtsai. (2004). Linguistic Markers of Psychological Change Surrounding September 11, 2011. *Psychological Science*, Vol 15, No 10, p. 687-693.

Chung, C. K. és Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*. Vol 42, No 1, p. 96–132.

Creswell, J. W. és Plano Clark, V. L. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks: SAGE Publications.

Currin-McCulloch, J. és mtsai. (2020). Understanding breast cancer survivors' information-seeking behaviours and overall experiences: A comparison of themes derived from social media posts and focus groups. *Psychology & Health*. Vol 36, No 7, p. 810-827.

Csepeli, Gy. (2015). A szociológia és a Big Data. *Replika*, Vol 2015, No 3-4, p. 169–174.

De Felice, R. és Garretson, G. (2018). Politeness at Work in the Clinton Email Corpus: A First Look at the Effects of Status and Gender. *Corpus Pragmatics.* Vol 2, p. 221–242.

Dessewffy, T. és Láng, L. (2015). Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. *Replika*. Vol 2015, No 3-4, p. 157–170.

Evans, J. A. és Aceves, P. (2016). Machine translation: mining text for social theory. *Annual Review of Sociology*, Vol 42, No 1, p. 21–50.

Furkó, B. P. (2019). Populist Discursive Strategies Surrounding the Immigration Quota Referendum in Hungary. In: Zienkowski, J. és Breeze, R. szerk., *Imagining the Peoples of Europe: Political Discourses Across the Political Spectrum*. Amsterdam, John Benjamins Publishing Company. p. 343–363.

Gödri, I. és Horváth, V. (2021). 'Nemzetközi vándorlás'. In: Monostori, J. és Őri, P. és Spéder Zs. szerk., *Demográfiai portré 2021*. Budapest, KSH Népességtudományi Kutatóintézet. p. 227–250.

Görög, R. (2017). Elvándorolt magyar véleményvezérek: Az online önreprezentáció és kapcsolatépítés lehetőségei a kivándorolt magyarok körében. *Médiakutató*, Vol 19, No 3/4, p. 119-133.

Grefenstette, G. (1999). 'Tokenization'. In: van Halteren, H. szerk., *Syntactic Wordclass Tagging.* Text, Speech and Language Technology, Vol 9. Dordrecht, Springer. p. 117-133.

Hajdu G. (2018). A kvantitatív és a kvalitatív társadalomtudományi kutatás módszerei-dióhéjban. *Forum Sententiarum Curiae*, Vol 2, p. 1–5.

Han, S. (2011). *Web 2.0*. New York: Routledge.

Blaskó, Zs. és Fazekas, K. szerk. (2016). *Munkaerőpiaci tükör, 2015.* Budapest: MTA Közgazdaság- és Regionális Tudományi Kutatóközpont Közgazdaság-tudományi Intézet.

Hárs, Á. (2020). 'Elvándorlás, visszavándorlás, bevándorlás. Jelenségek és munkaerőpiaci hatások.' In: Kolosi T., Szelényi I. és Tóth I. Gy. szerk., *Társadalmi Riport 2020*. Budapest, TÁRKI. p. 115-145.

Horváth, V. (2023). Transznacionális családi élet: A kiskorú gyermeket nevelő szülők családtól való elszakadással járó ausztriai munkavállalásának okai. *Társadalomtudományi Szemle*. Vol 13, No 3, p. 20–50.

Jun, S. P. és mtsai. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, Vol 130, p. 69-87.

Katona, E. és mtsai. (2021). Cikkismertetés: Depressziós fórumok témáinak automatizált szövegelemzése a depresszió biopszichoszociális modelljének tükrében. *Egészségfejlesztés*, Vol 62, No 4, p .76-79.

Katona, E. (2023). Természetesnyelv-feldolgozás a korrupciókutatásban: Új adatforrások, új módszerek, új tartalmi kérdések. *Socio.hu: Társadalomtudományi Szemle*. Vol 13, No 3. p. 76-95.

Kaur, J. és Buttar, P. K. (2018). A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, Vol 4, No 4, p. 207-210.

Kemp, S. (2023). Digital 2023: Global overview report. [Portál] DataReportal. Elérhető: https://datareportal.com/reports/digital-2023-global-overview-report. (Letöltve: 2024.03.28.)

King, G. (é.n.). Automated text analysis [szerzői honlap] Harvard University. Elérhető: https://gking.harvard.edu/category/research-interests/applications/automated-text-analysis (Letöltve: 2024.01.28.)

Király, G. és mtsai. (2014). Kevert módszertani megközelítések. Elméletek és módszertani alapok. *Kultúra és közösség*, Vol 5, No 2, p. 95-104.

Kmetty Z. (2018). A szociológia helye a Big Data-paradigmában és a Big Data helye a szociológiában. *Magyar Tudomány*, Vol 179, No 5, p. 683–692.

Központi Statisztikai Hivatal (é.n.[a]). [statisztikai adattábla] A népesség száma és átlagos életkora nem szerint. Elérhető: https://www.ksh.hu/stadat_files/nep/hu/nep0002.html (Letöltve: 2024.02.11.)

Központi Statisztikai Hivatal (é.n.[b]). [statisztikai adattábla] A kivándorló magyar állampolgárok célországok és nemek szerint. Elérhető: https://www.ksh.hu/stadat_files/nep/hu/nep0031.html (Letöltve: 2024.02.11.)

Landau, S. és Everitt, B. S. (2004). *A handbook of statistical analyses using SPSS.* Boca Raton: Chapman and Hall/CRC.

Markowitz, D. M. (2021). The meaning extraction method: An approach to evaluate content patterns from large-scale language data. *Frontiers in Communication*, Vol 6, p. 1-11.

Mayer-Schönberger, V. és Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Boston: Houghton Mifflin Harcourt.

Monzani, D, és mtsai. (2021). Emotional Tone, Analytical Thinking, and Somatosensory Processes of a Sample of Italian Tweets During the First Phases of the COVID-19 Pandemic: Observational Study. *Journal of Medical Internet Research,* Vol 23, No 10, p. 1-11.

Nagy Hesse-Bieber, S. (2010). Mixed Methods Research: Merging Theory with Practice. New York: Guilford Press.

Ophir, S. (2016). Big data for the humanities using Google Ngrams: Discovering hidden patterns of conceptual trends. *First Monday*. Vol 21, No 7.

O'Reilly, T. (2009). *What is web 2.0?* USA: O'Reilly Media.

Orosz, Gy. és mtsai. (2022). 'HuSpaCy: an industrial-strength Hungarian natural language processing toolkit'. In: Berend, G. és mtsai. szerk., *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem. p. 59–73.

Pham, M. D. és mtsai. (2023). What are we fighting for? Lay theories about the goals and motivations of anti-racism activism. *Race and Social Problems*. Advance online publication.

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Royal Statistical Society*. Vol 48, No 2, p 167–227.

Rhidenour K. B. és mtsai. (2019). Heroes or Health Victims?: Exploring How the Elite Media Frames Veterans on Veterans Day. *Health Communication*. Vol 34, No 4, p. 371-382.

Rodríguez-Arauz, G. és mtsai. (2017). Hablo Inglés y Español: Cultural Self-Schemas as a Function of Language. *Frontiers in Psychology*. Vol 8, p. 1-15.

Ross, A. S. és Rivers, D. J. (2018). Discursive Deflection: Accusation of "Fake News" and the Spread of Mis- and Disinformation in the Tweets of President Trump. *Social Media + Society*, Vol 4, No 2, p. 1-12.

Ságvári, B. (2017). Társadalomtudomány a Big Data korában. *Statisztikai Szemle*. Vol 95, No 5. p. 491–504.

Siklósi, B. és mtsai. (2016). Context-aware correction of spelling errors in Hungarian medical documents, *Computer Speech & Language*, Vol 35, p. 219-233.

Silva, J. X. és mtsai. (2023). ’Low-code and no-code technologies adoption: a gray literature review’. In: da Cunha, M. X. C. és mtsai. szerk., Proceedings of the XIX Brazilian Symposium on Information Systems. New York, Association for Computing Machinery, p. 388-395.

Simonovits, B. és Bernát, A. (2016). *The Social Aspects of the 2015 Migration Crisis in Hungary.* Budapest: TÁRKI.

Siskáné Szilasi, B. és mtsai. (2017). A magyar fiatalok erősödő kivándorlási szándékának kiváltó okai és jellemzői. *Tér és Társadalom.* Vol 31, No. 4, p. 131-147.

Szigeti, Á. (2022) A szövegbányászat mint társadalomkutatási módszer. *Szociológiai Szemle,* Vol 32, No 2, p. 91–100.

Wang, G. (2022). Britain as a protector, a mediator or an onlooker? Examining the 2019–20 Hong Kong protests in British newspapers. *Journal of Language and Politics*. Vol 21, No 1, p. 17-36.

Wilson, S. és mtsai. (2016). ’Disentangling Topic Models: A Cross-cultural Analysis of Personal Values through Words’. In: Bamnan, D. és mtsai. szerk., *Proceedings of the First Workshop on Natural Language Processing and Computational Social Science*. Austin, Association for Computational Linguistics. p. 143–15.

Zhang, Y. és Wildemuth, B. M. (2017). ’Qualitative Analysis of Content’. In: Wildemuth B. M. szerk*., Applications of Social Research Methods to Applications to Question in Information and Library Science*. Belmont, Brooks/Cole. p. 318-329.