A történelmi korszakhatárok mekkora változással vannak a szókincsükre, avagy a szavak gyakoriságának a vizsgálata

Bajkó Balázs (KJE), Pitlik László (MY-X team)

Kivonat: A cikk egy a Kodolányi János Egyetemen, szabadon válaszható kurzus keretein belül folyó munkálatok nyomán jött létre. A feladat célja, hogy a tanuló az oktatóval együttműködve véghez vigyen egy elemzési és publikációs folyamatot. A Hallgató által választott feladat témája annak vizsgálata, hogy a történelmi korszakhatárok mekkora változással vannak a szókincsükre, avagy a szavak gyakoriságának vizsgálata az Ngram viewer segítségével.

Ngram: Beszéd vagy egy szöveg tartalmának statisztikai vizsgálata, hogy feltárjuk a szövegben valamilyen elem (karaktersor) számát (beazonosítsa ennek a korpuszon belüli előfordulási arányát). Jelenleg az alapértelmezett érték 1800-2019 között van meghatározva, tehát az adatvagyon 220 év vizsgálatára képes. Különböző korpuszokban tudunk keresni, így lehetőségünk van többek között az Angol, Francia, Német, Olasz, Orosz, Spanyol, Kínai korpuszban vizsgálódni, amelyek nem csak nyelv szerint, de akár évek szerint is el vannak különítve egymástól.

A kutatás alapjául Minya Károly, Mai magyar nyelvújítás - Szókészletünk módosulása a neologizmusok tükrében a rendszerváltozástól az ezredfordulóig könyve szolgált (vö. 0. ábra):

A képen szöveg, képernyőkép, Betűtípus, dokumentum látható

Automatikusan generált leírás

1. Ábra: A szókészletváltozás potenciális szakirodalma (forrás: URL = …)🡨reklámsáv nélkül szebb lenne a képernyőkép!!! 😊

Vizsgált szavak: *democracy, colony, kingdom, peace , war*

A fenti angol szavakat a brit angol korpuszban vizsgáltam a fentebb jelzett 220 évre vonatkozóan.

A korrektúrák tudatosan kerültek meghagyásra a szövegben a Hallgató-oktató kooperáció rétegeinek, fokozatainak demontrálására.

# Bevezetés

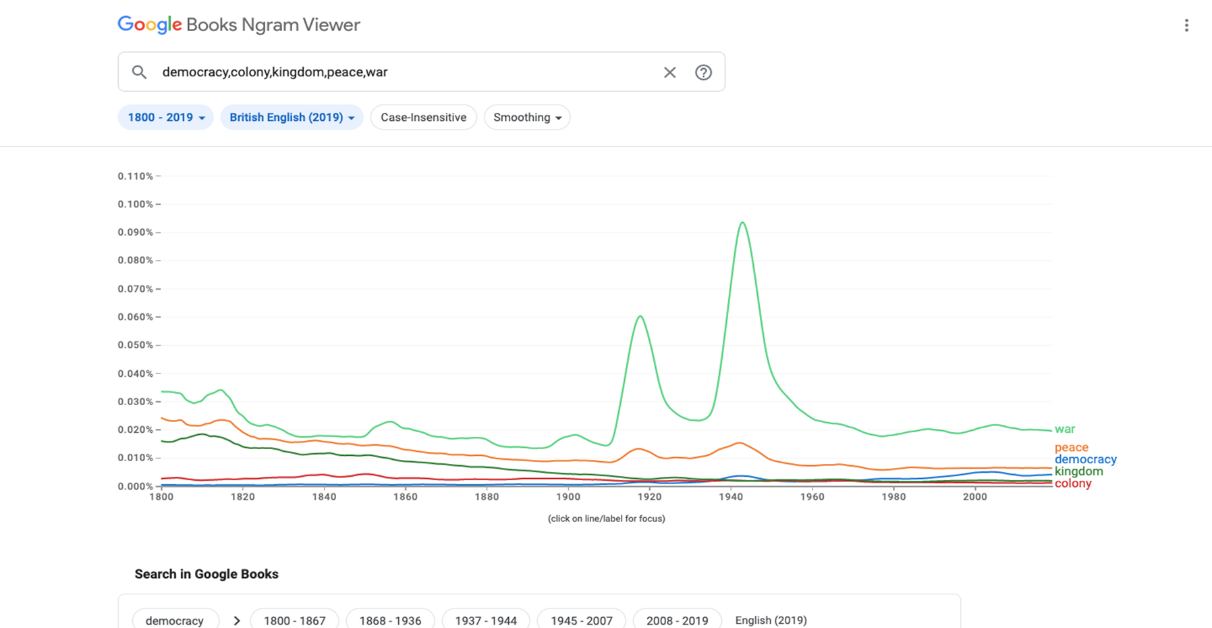
A feladat célja, hogy a tanuló az oktatóval együttműködve véghez vigyen egy adatelemzési, adatvizualizációs (pl. „képregény”-írási, elemzés-reprodukciós, tutoriál-szerkesztési, robot-fejlesztési) folyamatot, amelyet később akár publikálni is lehet.

Mit értünk nyelvtörténeti korszakhatár alatt?

Nyelvtörténeti korszakhatár alatt azt értem, amikor a történelem hatására bizonyos szavak addig nem, vagy csak nagyon alacsony számmal jelentek meg a korpuszban/ szókincsben. Jelen felmérésben azt szeretném bebizonyítani, hogy az első világháború nyelvtörténeti korszakhatárnak is tekinthető.

Mit értünk történelmi fordulópontként?

Történelmi fordulópontként az első világháborút jelöltem ki intuitív (naiv = nem optimalizált) alapon és az az előtti és utáni időszakát vizsgáltam a kiválasztott szavak segítségével a brit korpuszban. Az intuitív vélelem alapját az 1. ábra adja, ahol is a háború kifejezés karakteres (dátumhoz kötötten is értelmezhetőnek tűnő) eltéréseket (korszakhatárokat?) mutat fel a laikus szemlélő számára (is):



*1. ábra – A vizsgált szavak gyakorisága a brit korpuszban 1800-2019 között*

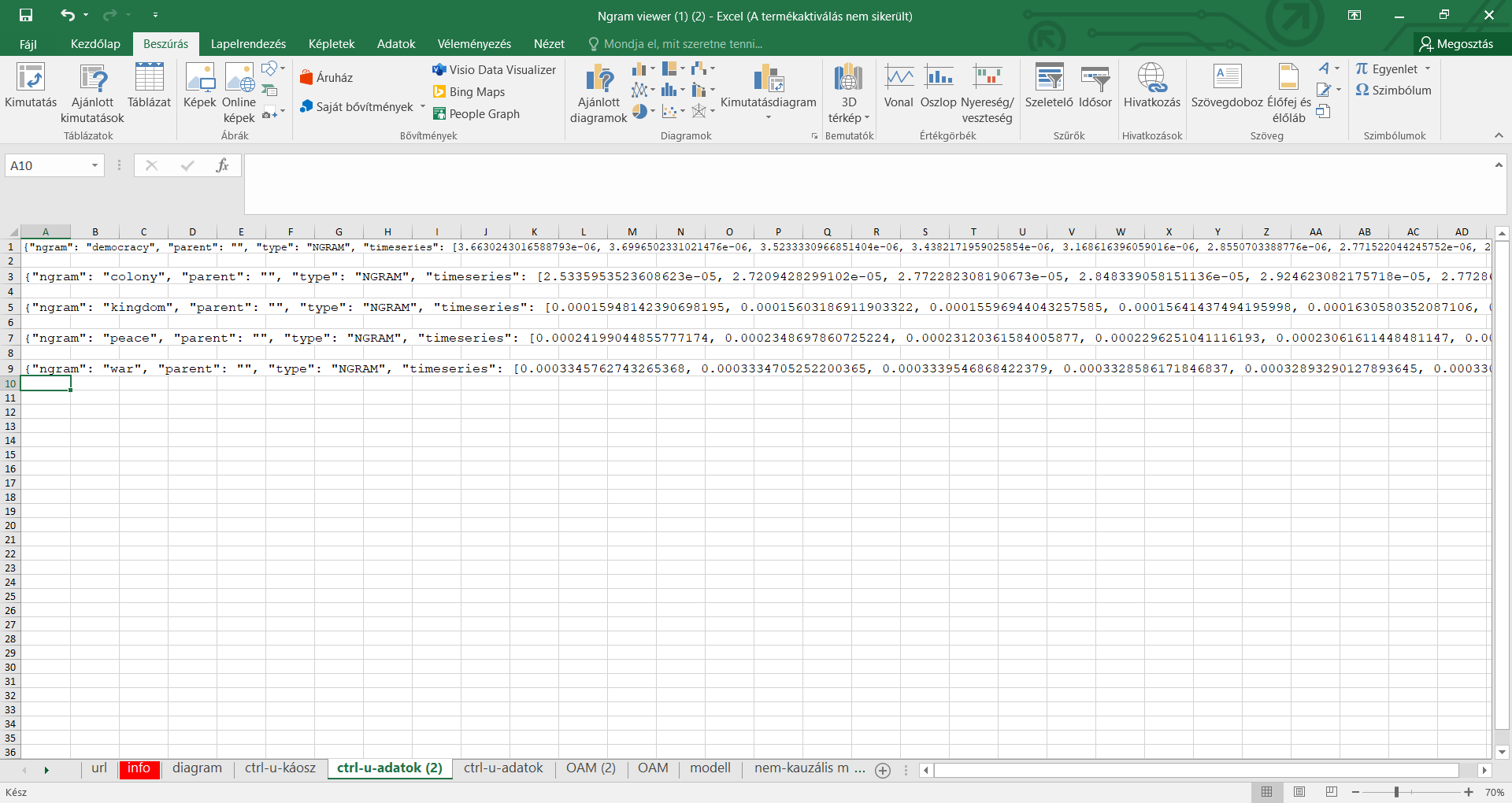
*Forrás: NGRAM–URL = ???*

*(ill.* <https://miau.my-x.hu/miau/301/Ngram%20viewer%20(1).xlsx> – diagram munkalap)

# Adatvagyon

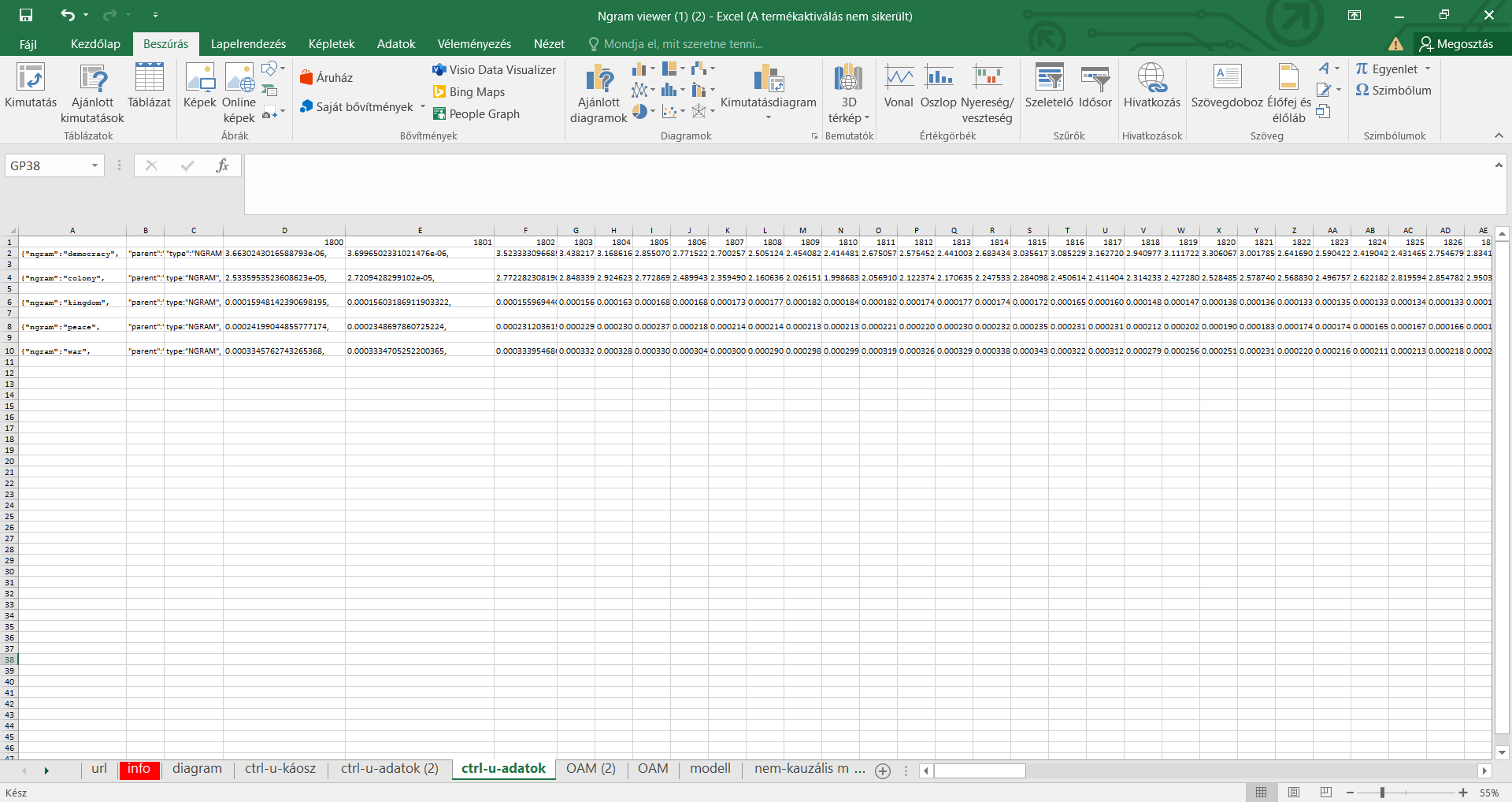
Háttér-XLS: URL = <https://miau.my-x.hu/miau/301/Ngram%20viewer%20(1).xlsx>

Az Ngram-ból CTRL+U billentyű kombinációval sikerült kinyernem a nyers adatokat. A „ctrl-u adatokban (2)” munkalapon sikerült az adatok strukturálisabb elhelyezése, és átláthatóbbá tétele, majd a „ctrl-u adatokban” munkalapon megkaptuk a feldolgozásra alkalmas nyers adatokat egy korábbi elemzés XLS-állományát klónozva, adaptálva: vö. https://miau.my-x.hu/miau/292/ngram.xlsx. A 2. és 3. ábra bemutatja az érdeklődő, reprodukciót/adaptációt tervező Olvasó számára a háttér-XLS-állomány kapcsolódó nézeteit:



*2. ábra – Nyers adatok az Ngram viewer adatbázisból a ctrl-u kombináció segítségével*

*Forrás: Saját kutatás -* <https://miau.my-x.hu/miau/301/Ngram%20viewer%20(1).xlsx>



*3. ábra- A kinyert adatok strukturálisabb formában*

*Forrás: Saját kutatás -* <https://miau.my-x.hu/miau/301/Ngram%20viewer%20(1).xlsx>

# Elemzések

Az öt használt szavam segítségével és Minya Károly elméletét alapjául véve be szerettem volna bizonyítani, hogy a történelem során bekövetkezett események hatására a szókincsünkben látható változás következett be. Ezért is próbáltam olyan szavakat vizsgálni, amelyek igazán gyakoriak és valószínűleg sokszor fordulnak elő (vö. 4. ábra – OAM munkalap – az 1. ábra ellenőrzéseként a CTRL+U alapú adatelőkészítési lépések nyomán saját adatvizualizáció keretében elkészítve a további adatfeldolgozási lépések alapjaként):

*4. ábra – Az adatok becslése a vizsgált szavakra lebontva🡨hol vannak a szavak???*

*Forrás: Saját ábra -* <https://miau.my-x.hu/miau/301/Ngram%20viewer%20(1).xlsx>

Akkor lesz igaza a szerzőnek, ha bebizonyítjuk, hogy a történelem változások során keletkezett szavak is beépültek szókincsünkbe és adott esetben egyre többször vagy egyre kevesebbszer kerül használatra a korpuszban.

A naiv "bizonyítás” maga a 4. ábra, ahol a „war” (felső sötétkék) kifejezés (vonal) látszólag karakteresen más, mint a többi. De ez a látszólagosság már pl. a felső világos sárga vonal kapcsán is értelmezhető, hiszen a legfelső sötétkék vonal elnyomja a következő (világosabb sárga) vonal hullámzásának karakterisztikus jellegét: vagyis már ez az alapkép is felhívja a figyelmet az arányok egymást torzítani képes mibenlétére!

Így a bizonyítás, mint olyat komplexebben kell definiálni: keressük tehát a tananyag ajánlásai értelmében azt, vajon minden év szóarányai tekinthetők egy a nyelvhasználat kapcsán azonos jellegűen (vö. minden út Rómába vezet?!).

Az 5. ábra (vö. OAM(2) munkalap) azt demonstrálja, milyen potenciális eredmény várható a fentebb definiált anti-diszkriminatív vizsgálattól:

*5. ábra- Az adatok becslése szavak összeségét vizsgálva*

*Forrás: Saját ábra -* <https://miau.my-x.hu/miau/301/Ngram%20viewer%20(1).xlsx>

Az anti-diszkriminatív vizsgálat eredménye egy „lázgörbe” vagy „EKG-görbe”, melynek X-tengelye értelemszerűen az idő (az évek, mint objektumok) és az Y-tengelye egy fiktív (mértékegység nélküli) hasonlósági skála, melynek centrális értéke egy véletlenszerűen választott konstans érték: itt és most 1.000.000 pont.

Minden objektum (év), amihez az optimalizáló eljárás (COCO Y0: https://miau.my-x.hu/my-x-factor-y) a centrális (norma) érték (1.000.000 pont) közelébe eső értékkel tér vissza, értelemszerűen norma-jellegű (semleges) évnek tekinthető a korszakhatárképzési szándékok szempontjából.

A korszakhatárt az EKG-görbéről le lehet olvasni szemmel, vagy az anti-diszkriminatív eljárás korszakhatár-kereső funkcióját erre a görbére önmagában is ki lehet terjeszteni: vö. <https://miau.my-x.hu/miau2009/index.php3?x=e0&string=cutting>, ill. <https://miau.my-x.hu/miau2009/index.php3?x=e0&string=szakaszolas>

Az 5. ábra születési körülményei rel. egyszerűek: Optimalizáció keretében keressük azon lépcsős függvény paramétereit, melyek minden egyes vizsgált szó (attribútum) minden egyes éve kapcsán helyettesítési értéket jelentenek, ahol a helyettesítési értékek szavanként kötelesek a kisebb nyersadat előfordulást kisebb csereértékkel büntetni és a teljes csereérték rendszer köteles minél közelebbi értékeket adni a centrális 1000000 pontos elváráshoz minden év esetén a csereértékek soronkénti összegeire vonatkozóan becslésként. A soronként becslések és a centrális érték eltéréseinek abszolút hibaösszege kerül minimalizálásra, ahol az 5. ábrán ábrázolt vonal elemei maguk ezen becslés értékei,

Érdemi korszakhatár-átcsapásról 1927 és 1936 között lehet beszélni, vagyis azok a figyelmes fülek/szemek, melyek az angol korpusz változásait pl. újságíróként, politikusként anno követték, a várható háborút (a 2. világháborút) már évekkel a tényleges megkezdés előtt kellett/lehetett, hogy érzékeljék. Ezt meg is tették vélhetően sokan, hiszen pl. az egyre fasizálódó Németországból való elmenekülés éppen ebben az időszakban volt még lehetséges…

Az, hogy az első világháború környékén nem látszik hasonló átcsapás, ill. az első világháború előtti közel normálisnak érzékelhető kilengéshez képest egy hullámzó süllyedés látható, itt és most önkritikusan jelezve az elemzés és a szómágia kapcsolatát, vagyis némi belemagyarázással úgy fogalmazható meg, hogy az első világháború nem oldott meg semmit, sőt, még több bajhoz vezetett „béke”-kötési feltételeivel, mint ami korábban már volt.

Nem mellesleg 1848 előtt/táján szintén látható egy „szóhasználati földrengés”, melynek eredményei a az egyes országok forradalmi helyzeteiben érhetők tetten.

A Szovjetunió szétesését egy, az olajválság (1973) időszakától 1995-ig húzódó szóhasználati szeizmikus mintázat vezeti fel/le…

Mindenképpen vészjelzésként értelmezendő, hogy a 2008/2009-es „pénzügyi válság” óta a szóhasználati földrengések mintázatai nem hasonlítanak egy kicsit sem a 2. világháború utáni, vagy a Szovjetunió szétesése utáni rel. nyugalomhoz…

# Eredmények

Az 5. ábra tehát már naivan is jól szemlélteti egy történelmi fordulópontok hatásait, illetve magát a történelmi fordulópontot és a szavakra gyakorolt hatását. Az látni lehet, hogy minden nagyobb történelmi eseménynél a grafikon kiugrást mutat: ahogy a diagramm is mutatja például az 1848-ban a hazánkat is elért európai forradalmi hullám időszakában a grafikon felívelést mutat, az azt megelőző időszakhoz képest. Nem volt ez másként az Első világháború időszaka alatt is tapasztalható ugyanez a jelenség, illetve az egyik legnagyobb kiugrás a Második világháborút megelőző 6-8 évben a legszembetűnőbb (vö. militáns politikai szándékokat előkészítő propaganda gépezetek előkészítő működése?!), itt egy igen magas kiugrásról beszélhetünk, ami nem volt jellemző az ezt megelőző vizsgált időszakokhoz képest. A Második világháború után kisebb-nagyobb kilengések figyelhetőek meg, majd napjainkban a 2008-as Gazdasági világválság idején, illetve 2014 környékén láthatunk hasonló jelenségeket.

# Vita

A Háttér-XLS OAM munkalapja tartalmaz egy másik elemzéshez kapcsolódó lázgörbét is. Míg az OAM munkalapon az I1:M1 tartomány végig nulla értékeket tartalmaz az vizsgált szavak kapcsán, addig az OAM(2) esetén ezek a paraméterek:



A paraméterek jelentése: 0 = minél nagyobb, annál nagyobb legyen a centrális értékre vonatkozó becslés, 1 = minél kisebb az input nyers előfordulási gyakorisága, annál nagyobb legyen a becslés

A két modell értelmezési különbsége: A végig nulla paraméter-értéksor jelentése: bármiről is, de fecsegjenek az emberek. A vegyes paramétersor jelentése: mekkora a békevágy? Vagyis az 5 ábra nem egyszerűen csak egy EKG, hanem a békevágy alakulásának leképeződése!

6. ábra: Fecsegés-index (forrás: saját ábrázolás, háttér-xls, oam munkalap)

A 6. ábra jelentése: a lázgörbe, a fecsegés index trendje csökkenő, vagyis a háború és béke svédasztala az embereket egyre kevésbé érdekli (vö. Harari: ilyen jól még sosem élt az emberiség vélelme?! - https://www.google.com/search?q=harari+ilyen+j%C3%B3l+m%C3%A9g+sosem+%C3%A9lt+az+emberis%C3%A9g)…

A 2012-es rel. mély beszakadás pontos oka talán még nem is került szómágikusan értelmezésre (vö. <https://hu.wikipedia.org/wiki/2012> – 200+ nyelven más és más súlypontokkal). De ez az év a békevágy történelmi minimumát súrolja!

# Konklúzió

Az Ngram egy nagyszerű lehetőség volt az elemzést megalapozó hipotézis kialakítására. A lekért adatok egy átfogó képet adtak arról, amit vizsgálni szerettem volna, így további vizsgálatokra (látszólag) nincs szükség. Nyugodt szívvel kijelenthető, hogy a történelem teljes mértékben hatással van a szókincsünkre és a fordulópontnak nevezhető események után kiugrást mutatnak igen sok esetben. Az már csak az olvasón múlik, hogy mit lát bele, vagy esetlegesen mivel tudja magyarázni ezeket a jelenségeket. Pont emiatt lesz ez egyfajta „jóslás” (vö. MI-alapú asztrológia /a Nostradamus-gép, stb.). Hasonló kihívások már korábban is kezelésre kerültek (vö. https://miau.my-x.hu/miau2009/index.php3?x=e0&string=golden)...

# Források

<https://hu.eyewated.com/a-ngram-viewer-eszkoez-hasznalata-a-google-koenyvekben/>

<https://books.google.com/ngrams/>

<http://www.tintakiado.hu/book_view.php?id=86>

A korrektúrák láthatóként való meghagyása tudatos szerkesztési döntés a Hallgató-oktató-együttműködések szintjeinek, rétegeinek demonstrálása céljából…