CIA szövegek kohéziójának elemzése első- és másodrendű Markov láncokkal

Cohesion analysis of the CIA texts based on first and second order Markov Chains

Dr. TÓTH Erzsébet[1], Dr. GÁL Zoltán[1]

[1]Debreceni Egyetem Informatikai Kar
4028 Debrecen, Kassai út 26., Magyarország, http://www.inf.unideb.hu

Abstract

The aim of this paper is to quantitatively analyze the cohesion of political and military texts from the Central Intelligence Agency (CIA) digital library using Markov chains. The texts were characterized by 17-dimensional feature vectors (FVs), which describe the ratio of parts of speech. The texts were divided into 80 text entities, the similarity between each entity and the entire text was calculated using cosine distance, and from these we generated time series describing the dynamics of the text. The cohesion of the texts is represented by the overall thematic cohesion (OTC) measure. According to our results, the 30 CIA texts examined show high cohesion, which increases with the length of the text, and the application of a second-order Markov model is more sensitive to the temporal structure of the text dynamics.

Keywords: text cohesion metrics, 1st and 2nd order Markov chains, DBSCAN algorithm, Linkage algorithm, k-means algorithm.

Kivonat

A dolgozat célja a Central Intelligence Agency (CIA) digitális könyvtárából származó, politikai és katonai témájú szövegek kohéziójának kvantitatív elemzése Markov-láncok alkalmazásával. A szövegeket 17 dimenziós tulajdonság vektorokkal (FV) jellemeztük, amelyek a szófajok arányát írják le. A szövegeket 80 szövegentitásra bontva, az egyes entitások és a teljes szöveg közötti hasonlóságot koszinusz-távolság segítségével számítottuk ki, és ezekből állítottuk elő a szöveg dinamikáját leíró idősorokat. A szövegek kohézióját az átfogó tematikus kohézió (OTC) mérőszám reprezentálja. Eredményeink szerint a vizsgált 30 CIA-szöveg magas kohéziót mutat, amely a szöveghosszal növekszik, és másodrendű Markov-modell alkalmazása érzékenyebben ismeri fel a szövegdinamika időbeli szerkezetét.

Kulcsszavak: a szövegkohézió metrikái, első és másodrendű Markov láncok, DBSCAN algoritmus, Linkage algoritmus, k-Means algoritmus.

1. BEVEZETÉS

A kvantitatív nyelvészet területe számos törvényt és mérőszámot foglal magába a szöveges tartalmakban előforduló különböző mintázatok elemzésére. Dolgozatunkban a szövegek kohéziójának és azok dinamikájának leírására releváns statisztikai metrikákat és matematikai modelleket alkalmazunk. Szövegkohézió fogalma alatt a szöveg különböző részeit összetartó erőt értjük, amely biztosítja, hogy a mondatok és a gondolatok logikusan kapcsolódjanak egymáshoz, jól érthető egészet alkotva. Ezt nyelvtani (lineáris) és tartalmi (globális) eszközökkel érik el, mint például kötőszavak, ismétlés, névmások, ill. a téma-réma szerkezet vagy a logikus gondolatmenet. Korábbi kutatásainkban [1] megállapítottuk, hogy egy adott szövegen belül a különböző entitások (azaz szövegrészletek) és a szülő szöveg között magas szintű korreláció figyelhető meg, ami az adott szöveg kohéziójával magyarázható. Ugyanez a magas szintű korreláció nem áll fenn a különböző szövegek eltérő entitásai között. A narratív folyamat az a fontos fogalom, amin keresztül a szerző megvalósítja céljait, ezt tanulmányozva írhatjuk le, hogy hogyan működik a narratíva. A szövegdinamika a narratíva belső folyamata, ami által eléri céljait, az olvasói dinamika az olvasó ezzel

megegyező kognitív, etikai, affektív és esztétikai válasza a szövegdinamikára [6]. Vizsgálatunkban szövegdinamika fogalma alatt az adott szövegben található események, történések sorát értjük, amit a szerző elmond a hallgatóságának. A szövegek láncszerű szerkezetűek, folyamatosan előrehaladva bontakozik ki bennük a cselekmény. A cselekmény közlése történhet különböző idősíkok között váltva, valamint az események, a dolgok közepébe vágva, azaz „in medias res".

Kapcsolódó kutatásnak tekinthetjük az ókori görög irodalmi szövegek osztályokba történő sorolását, amit Visszacsatolásos Neurális Hálózat (RNN) alkalmazásával hajtottunk végre az ókori Alexandriai Könyvtár osztályozási rendszerére támaszkodva [2, 3]. Egy másik irányban is kutatást végeztünk, ahol a MARCELL projekt során összegyűjtött Európai Uniós horvát és angol nyelvű jogszabályokat elemeztük. Elvégeztük a jogi korpuszban lévő címkézetlen jogi szövegek téma besorolásának becslését a Latent Dirichlet Allocation (LDA) algoritmus alapján több címkés osztályozást használva. Kidolgoztuk az LDA módszer flexibilis változatát, hogy azzal támogassuk a jogi szövegek minél kifinomultabb téma besorolását, ahol adott küszöbérték definiálásával több dobogós téma besorolását valósítottuk meg a jogszabályok számára [4, 5].

A dolgozat további részének felépítése a következő: az elsőrendű és másodrendű Markov láncok modellezési tulajdonságainak bemutatása a második részben található. Utána a CIA (Central Intelligence Agency) könyvtárából letöltött szövegek elemzése kerül sorra Markov módszerrel. Az összefoglalás és jelen kutatási tevékenység folytatásának lehetséges lépései a dolgozat végére került.

2. ALKALMAZOTT MÓDSZER

Az elemzett szövegek szerzői események, történések sorát írják le adott időrendi sorrendben. A szöveg szerzői saját egyéni stílusukat megvalósítva mondják el a történetet olvasóiknak. A szerzők által megtervezett és felépített szöveg annak kohézióját eredményezi, ami kiaknázható minden egyes szöveghez tartozó szövegrészlet automatikus leírására. A dolgozatban szövegentitásnak tekintjük adott szöveg egymás után következő mondataiból álló szövegrészletet. Ebből a szempontból nézve tehát a szöveg szövegentitások sorozatának fogható fel. A szövegek kohéziójának és dinamikájának leírására első- és másodrendű Markov láncokat alkalmaztunk matematikai modellként.

A diszkrét idejű Markov lánc az $\{Xt\}, t \geq 0$, sztochasztikus idősorok modellezéséhez használt módszer, amelyeknél a következő megfigyelés csak korlátozott számú előző megfigyeléstől függ:

$$P(X_{t+1} = x_{t+1} | X_t = x_t, \dots, X_0 = x_0) = P(X_{t+1} = x_{t+1} | X_t = x_t, \dots, X_{t-k} = x_{t-k}) \quad (1)$$

ahol $P(A|B)$ az A esemény B eseménytől függő feltételes valószínűségét jelenti. A következő esemény k darab előző eseménytől függ, ahol k-t a Markov folyamat rendjének nevezzük. Ha $k = 1$, akkor a folyamatot egyszerűen *(elsőrendű) Markov* folyamatnak, ha $k = 2$, akkor *másodrendű Markov* folyamatnak nevezzük. Megfigyelhető, hogy $k = 1$ esetén a jövő csak a jelentől függ, míg $k = 2$ esetén a jövő a múlttól és a jelentől is függ. Ezeknél a folyamatoknál az állapotok közötti váltások valószínűségét *P* tranzíciós mátrixba rendezzük. Adott sor-oszlop által jelölt elem a sor azonosítójú állapotból az oszlop azonosítójú állapotba való váltáshoz tartozik:

$$P_{ij} = P\left(X_{t+1} = s_j \middle| X_t = s_i, \dots, X_{t-k} = s_k\right), \text{ ahol} \quad (2)$$

$$\sum_{j=1}^{n} P_{ij} = 1 \quad (3)$$

vagyis a tranzíciós mátrix minden sora mentén teljes eseményhalmazunk van, azaz a valószínűség értékek összege 1). Az elsőrendű Markov folyamatot emlékezet nélkülinek nevezzük és jól alkalmazható a rövid távú idősoros függések modellezésére. A másodrendű Markov folyamat három dimenziós tranzíciós tenzorját is lehetséges kibővített mátrixra konvertálni:

$$P_{ijk} = P\left(X_{t+1} = s_k \middle| X_t = s_j, X_{t-1} = s_i\right) \quad (4)$$

$$P_{(ij)\to(jk)} = P\left(X_{t+1} = s_k \middle| X_t = s_j, X_{t-1} = s_i\right) \quad (5)$$

Itt a következő állapot az előző kettőtől függ, így erősebb időbeli szerkezetet képes megragadni. Az állapotpárok egyetlen kibővített állapotként való kezelése elsőrendű lánccá alakítja a folyamatot az $\mathbb{S}^2$ halmazon. Ez akkor hasznos, ha az egylépéses memória nem elegendő az idősor dinamikájának modellezésére. Egy diszkrét idejű Markov-lánc $\{Xt\}, t \geq 0$, stacionárius eloszlása $\boldsymbol{\pi} = (\boldsymbol{\pi_1}, \dots, \boldsymbol{\pi_n})$ egy olyan valószínűség-eloszlás, amely kielégíti az alábbi feltételeket:

$$\pi P = \pi \quad (6)$$

$$\sum_{i=1}^{n} \pi_i = 1 \tag{7}$$

ahol P az $n \times n$-es tranzíciós mátrix, és $\pi_i = P(X_t = s_i)$ az i-edik állapot stacionárius valószínűsége. Ha a lánc eléri a stacionárius eloszlást, akkor minden későbbi lépésben az eloszlás változatlan marad:

$$\lim_{t \to \infty} \mu_0 P^t = \pi \tag{8}$$

A (6) összefüggés alapján látható, hogy a π stacionárius eloszlás a tranzíciós mátrix saját vektora. A (7) tulajdonság alapján a P mátrix legnagyobb abszolútértékű sajátértéke $|\lambda_1| = 1$. A stacionárius eloszlás elérésének sebességét a P második legnagyobb sajátértékkel lehet mérni az alábbi módon:

$$\tau_{mix} = \frac{1}{1-|\lambda_2|} \tag{9}$$

ahol a nevezőben lévő $1 - |\lambda_2|$ tag a spektrális rés nevet viseli. Ez minél nagyobb, annál gyorsabban keveredik a lánc, azaz annál hamarabb közelít a stacionárius eloszláshoz. Ezt a metrikát használjuk a szövegek alszövegeinek elemzéséhez [7, 8, 9, 10].

3. ESETTANULMÁNY

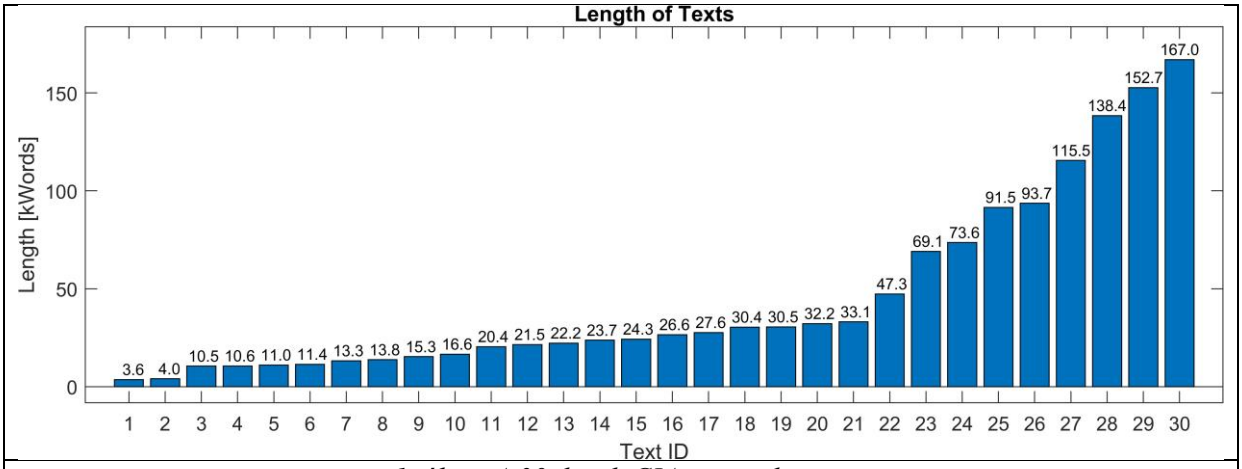
Korábbi vizsgálatunkhoz képest további 10 darab új szöveget töltöttünk le a Central Intelligence Agency (CIA) szervezet Digitális Könyvtárából, ami összesen 30 darab szövegből álló adathalmazt eredményezett. A szövegekről elmondható, hogy különféle politikai és katonai eseményeket írnak le, amelyek a világban történtek 50 évvel ezelőtt. Ezért ezek a szövegek homogénnek tekinthetők, mert a fent említett témákról szólnak. Ezen vizsgált szövegek felsorolását lásd az 1. táblázatban, ahol a szövegek a szavak számának növekvő sorrendjében szerepelnek.

Az elemzett CIA szövegek 1. táblázat

	Szerző(k)	Cím
1.	Central Intelligence Agency	Memorial Wall Publication
2.	Richard Mobley, James Marchio, Gary B. Keeley	The Thrill of the Hunt: Lessons from Archival Research into Cold-War Era Intelligence Decisionmaking
3.	Central Intelligence Agency	Notes from Our Attic: A Curator's Pocket History of the CIA
4.	Central Intelligence Agency	The Work of a Nation: The Center of Intelligence
5.	Central Intelligence Agency	Our First Line of Defense: Presidential Reflections on US Intelligence
6.	Gregory F. Treverton, Renanah Miles	Unheeded Warning of War: Why Policymakers Ignored the 1990 Yugoslavia Estimate
7.	US Government	A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis
8.	David W. Waltrop	An Underwater Ice Station Zebra
9.	Central Intelligence Agency	The Caesar, Polo and Esau Paper: Cold War Era Hard Target Analysis of Soviet and Chinese Policy and Decision Making 1953-1973
10.	Central Intelligence Agency	A Life in Intelligence – The Richard Helms Collection
11.	Central Intelligence Agency	At Cold War's End
12.	Clayton D. Laurie, Andres Vaart	CIA and the Wars in Southeast-Asia 1947-75
13.	Central Intelligence Agency	Bosnia, Intelligence, and the Clinton Presidency: The Role of Intelligence and Political Leadership in Ending the Bosnian War
14.	Central Intelligence Agency	Penetrating the Iron Curtain: Resolving the Missile Gap with Technology
15.	Michael Warner, J. Kenneth McDonald	US Intelligence Community Reform Studies since 1947
16.	Central Intelligence Agency	President Nixon and the Role of Intelligence in the 1973 Arab-Israeli-War
17.	Central Intelligence Agency	The Warsaw Pact: Treaty Friendship, Cooperation and Mutual Assistance
18.	Central Intelligence Agency	Profiles in Leadership: Directors of the Central Intelligence Agency and Its Predecessors 1941-2023
19.	Central Intelligence Agency	Ronald Reagan: Intelligence and the End of the Cold-War
20.	Central Intelligence Agency	President Carter and the Role of Intelligence in the Camp David Accords

21.	Central Intelligence Agency	Predicting the Soviet Invasion of Afghanistan: The Intelligence Community's Record
22.	Andrew Skitt Gilmour	A Middle-East Primed for New Thinking: Insights and Policy Options From the Ancient World
23.	Robert Vickers	The History of CIA's Office of Strategic-Research, 1967-81
24.	Thomas L. Ahern, Jr.	„Nothing if Not Eventful": Recollections of a Life's Journey in CIA
25.	James W. „Bill" Lair as told to Thomas L. Ahern, Jr.	„An Excellent Idea: Leading CIA Surrogate Warfare in Southeast Asia, 1951-1970, a Personal Account
26.	Central Intelligence Agency	CIA's Analysis of the Soviet Union 1947-1991: A Documentary Collection
27.	John L. Helgerson	Getting to Know the President: Fourth Edition: Intelligence Briefings of Presidential Candidates and Presidents-elect, 1952-2016
28.	Central Intelligence Agency	Watching the Bear: Essays on CIA's Analysis of the Soviet Union
29.	Douglas F. Garthoff	Directors of Central Intelligence as Leaders of the U.S. Intelligence Community, 1946-2005
30.	L. Britt Snider	The Agency and the Hill: CIA's Relationship with Congress, 1946-2004

A szövegek hossza 3,6 és 167 ezer szó között található. A vizsgált szövegek 70%-a kevesebb, mint 35000 szót foglal magába, ezáltal lehetővé válik számunkra a szövegméretek megközelítőlegesen három lineáris értéktartományát vizsgálnunk (lásd az 1. ábrát).



1. ábra. A 30 darab CIA szöveg hossza

Mindegyik CIA szöveget tokenekre (szózsetonokra) alakítottuk át. Ezután a szövegek tokenjeit megfelelő szófaj kategóriába beazonosítottuk, különös tekintettel az adott szöveg kontextusára. Matlab programozási eszköz beépített függvényei segítségével megkaptuk az egyes tokenek megfelelő szófaj kategóriába történő besorolását.

A feature vektor token kategóriái *2. táblázat*

ID	Token kategória
1	melléknév
2	értelmező
3	határozószó
4	segédige
5.	kötőszó

ID	Token kategória
6	elválasztószó
7	indulatszó
8	főnév
9	számnév
10	egyéb

ID	Token kategória
11	viszonyszó
12.	névmás
13	tulajdonnév
14	írásjel
15	alárendelőszó

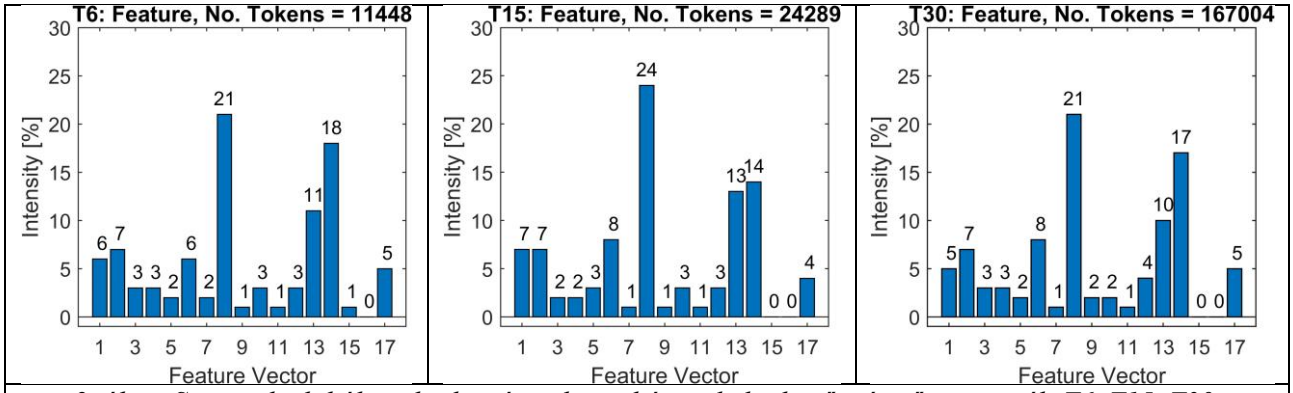
ID	Token kategória
16	szimbólum
17	ige

Összesen 17 token kategóriát használtunk (lásd a 2. táblázatot). Az „egyéb" token kategóriára azért volt szükség, mert a nemzeti nyelvek lehetséges speciális tulajdonságai ebbe helyezhetők el. Az egyes dimenziókban az adott szófajok százalékos előfordulási gyakorisága jelenik meg. Ez a 17 dimenziós jellemzővektor minden szövegegységben számszerűsíti az egyes szófajok arányát, gyakoriságát és eloszlását.

Lehetővé teszi a szövegek nyelvtani szerkezetének, stílusának és komplexitásának objektív összehasonlítását, valamint rejtett mintázatok és szerzői sajátosságok feltárását.

3.1. Szövegek tulajdonság vektorainak egyedi mintázatai

A vizsgált szövegeket egy 17 dimenziós FV (Feature Vector) tulajdonság vektor jellemzi a tokenek relatív száma alapján. Ezek a vektorok speciális mintázatukból adódóan tulajdonképpen leírják az adott szöveg jellegét a szövegek méretétől függetlenül (lásd a 2. ábrát).



2. ábra. Szövegek globális tulajdonság vektora három különböző méretű szövegnél: T6, T15, T30.

Megfigyelhető, hogy az azonos témákba tartozó szövegek is eltérő tulajdonság vektor mintázattal rendelkeznek a nyelv specifikussága, a szöveg stílusa, illetve másik szerzője miatt. Három kiválasztott szöveg eltérő tulajdonság vektor mintázata kerül bemutatásra a 2. ábrán. Az intenzitás százalékértékeket lefelé kerekítettük a [0, 100] intervallumban. Szembetűnő a diagramokon, hogy a legnagyobb intenzitás értékkel a főnév token kategória rendelkezik mindhárom esetben. A legritkább token kategóriák pedig a számnév, viszonyszó, az alárendelőszó és a szimbólum.

A $N = 30$ darab szöveg mindegyik ST_i^j szövegentitására (alszöveg, „SubText") vonatkozóan előállítottuk a $FV_i^j \in M_{1x17}$ tulajdonság vektorokat, ahol $i = 1, \dots, N$ és $j = 1, \dots, m$. A szövegentitások száma $m \in \mathbb{N}$, jelen esetben $m = 80$ darabra osztottuk fel mindegyik szöveget. Mivel a szövegek különböző hosszúságúak, ebből adódóan a különböző szövegekből származó szövegentitások is változó hosszúságúak (lásd az 1. ábrát).

3.2. Szövegentitások hasonlóságának elemzése

Minden egyes T_i, $i = 1, \dots, N$ szöveg esetén az $j = 1, \dots, m$ darab ST_i^j szövegentitására meghatároztuk az m darab FV_i^j tulajdonságvektort, illetve a szöveg egészére vonatkozó FV_i, globális tulajdonságvektort is. Ezáltal T_i esetén kiszámolható az m darab saját szövegentitás, illetve a globális tulajdonságvektor közötti hasonlóság. Ehhez a Koszinusz távolság metrikát használtuk, mivel az egyes FV_i^j vektorok egyes dimenziókban nagyon eltérő értékeket mutattak. A Koszinusz távolság kiszámolása két $u, v \in M_{1x17}$ vektor esetén az alábbi összefüggéssel történik:

$$\mathrm{d}(u, v) = 1 - \frac{u \cdot v}{\|u\|\|v\|} \tag{10}$$

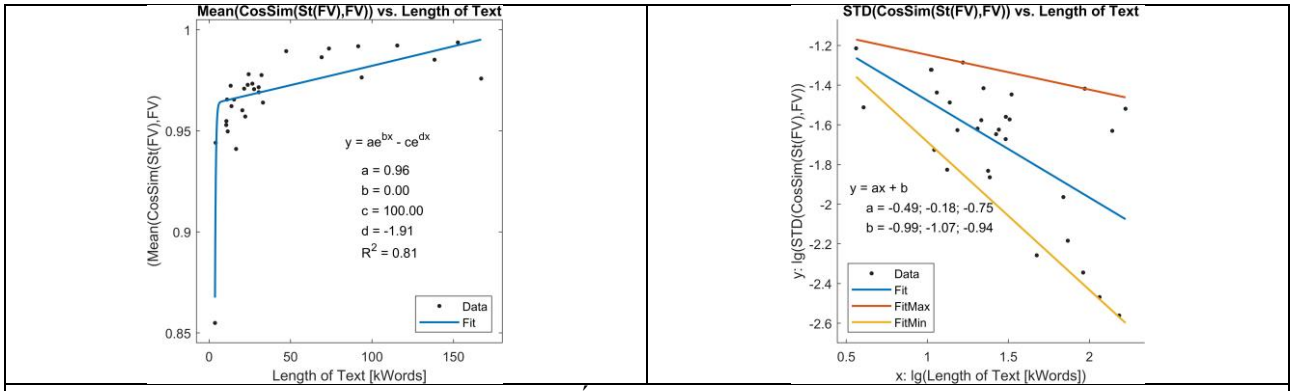
ahol $u \cdot v$ az két vektor belső szorzatát jelenti. Fontos megjegyezni, hogy a vektorok hosszától független ez a távolság, csupán a köztük lévő szöget érzékeli. Így párhuzamos vektoroknál a metrika nulla értéket ad, míg merőlegesség esetén az érték 1. Ennek segítségével az adott T_i szöveget egy $D_i = (d_i^1, \dots, d_i^m)$ idősorral jellemezzük, amit *szöveg dinamikájának* nevezünk. A D_i idősor átlaga az adott T_i szöveg egydimenziós értékét jelenti, vagyis a szöveg egy globális metrikája. Rögzített i esetén a d_i^j, j = 1,…,m hasonlóságok mindegyike az adott szöveg tartalmára jellemző.

Mivel az egymás után következő szövegentitások mondanivaló szempontjából összefüggők, ezért mindegyikük hozzájárul a teljes szöveg tartalmi integritásához, amit kohéziójának nevezünk. Ez alapján adott T_i szöveg esetén (i rögzített) a fentebb számolt D_i vektor átlaga a teljes szöveg tartalmára vonatkozó mérőszám, amit a szöveg OTC (Overall Thematic Cohesion), á*tfogó tematikus kohéziójának* nevezünk:

$$OTC_i = \frac{1}{m}\sum_{j=1}^{m} d_i^j, \quad i = 1, \dots, m \tag{11}$$

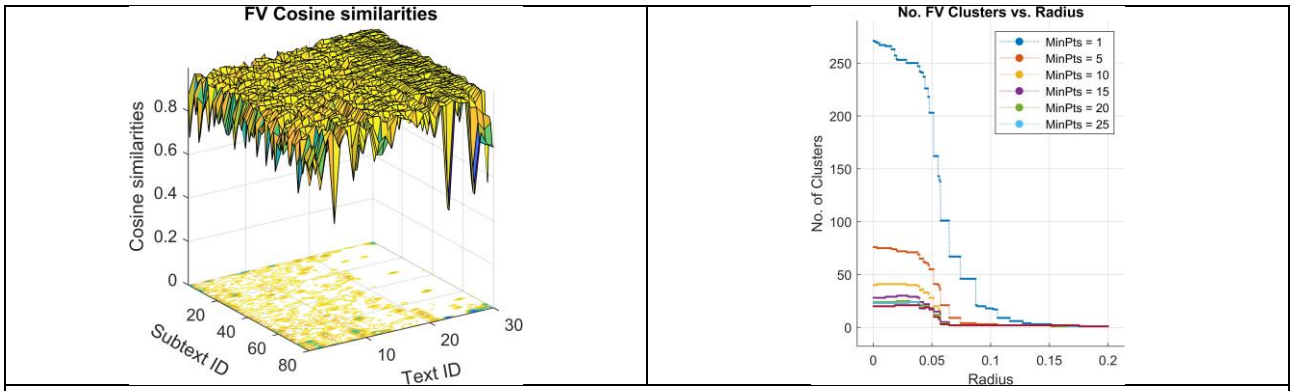
Könnyen belátható, az OTC metrika a $(0, 1)$ intervallumban adja értékeit. Az $N = 30$ szöveg esetén az átfogó tematikus kohézió mérettől függő tulajdonságát a 3. ábra szemlélteti. Megállapítható, hogy a vizsgált CIA szövegek esetén az OTC metrika minden esetben 85% feletti, tehát jól méri a szöveg tartalmi összetartozását. Ugyanakkor az is megfigyelhető, hogy az OTC a szövegek szavainak számától is függ. Minél hosszabb a koherens szöveg, annál nagyobb az OTC értéke is a $(0, 1)$ intervallumban. Ezt várhatóan az okozza, hogy a szerző több stilisztikai eszközt tud használni a nagyobb szövegnél a szövegentitások összekötéséhez.

A D_i idősorok szórása a szöveg *tartalmi témaváltásának* mértékét méri. Az N darab elemzett CIA szöveg esetén ezek a témaváltások egy hatványfüggvénnyel közelíthetők. Ezt szemlélteti a 3. ábra jobb oldala, ahol a log-log skálán egyenesek által határolt területünk van. Itt is megfigyelhető, hogy a szórás mértéke a szöveg szavakban kifejezett számától függ. Hosszabb szövegek szórása erőteljesebben eltér egymástól, mint a rövidebb szövegeknél. Ezt okozhatja a szókincs változatossága az egyes szerzők szövegében.



3. ábra. A szövegek kohézió metrikái: bal) Átfogó tematikus kohézió; jobb) Témaváltás intenzitás.

A T_i szövegnek a D_i idősor által képviselt dinamikája az elemzett CIA szövegek esetén zömében 70% felettiek (ld. 4. ábra bal oldal). A D_i mintázatok kisebb szövegnél nagyobb kilengéseket tartalmaznak, míg hosszabb szövegek esetén ez kevésbé változik adott szövegen belül. Ezt látjuk a 4. ábra bal oldalán lévő vetületeknél is. Ennek oka a szövegentitások méretéből adódó erősebb korreláció rögzített i esetén az FV_i^j és az FV_i között $j = 1, \dots, m$ értékekre. Ez a kohézió növekedését is okozza a mérettel.

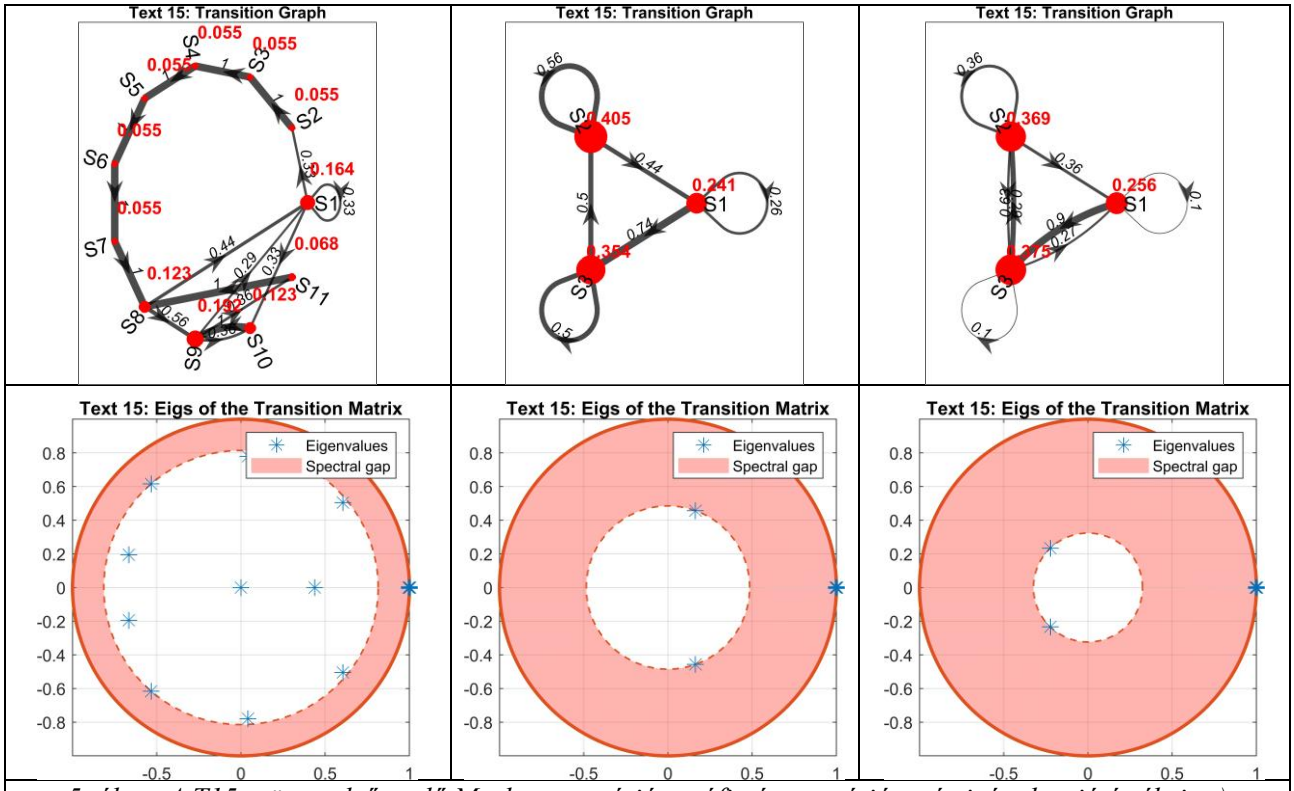


4. ábra. A szövegek hasonlóság metrikája: bal) Koszinusz távolság az alszöveg- és a szövegazonosító függvényében; jobb) A szövegek klaszter számának függése a sugártól.

Adott T_i szöveg $FV_i^j, j = 1, \dots, m$ hasonlóság jellemzőinek klaszterbe sorolásához DBScan algoritmust is használtunk. Mivel azonban a DBScan erőteljesen függ az $(r, MinPts)$ klasztersugár, illetve minimális elemszám paraméterektől, ezért több próbálkozással határoztuk meg ezek optimális értékét (ld. 4. ábra jobb oldal). Ha a klaszter tagjainak minimális számát (MinPts) kicsire vesszük, akkor sok kis csoport alakítható ki akkor is, ha a klaszter sugara kicsi. Ha a klaszterek minimális sugarát növeljük, akkor egyre több elem sorolható be ugyanabba klaszterbe, így a csoportok száma egyre csökken, mígnem az összes elem egyetlen közös csoportba sorolódik. A két paraméter helyes értékének meghatározása DBScan esetén úgy történik, hogy a néhány tagszám előírás mellett a csoportok számának sugártól való függésnél a legmeredekebb csökkenés közepét választjuk. Ez alapján az alkalmazott kombináció: $(r, MinPts) = (5\%, 5)$.

3.3. Első- és másodrendű Markov láncok alkalmazása

Három különböző klaszterezési algoritmust használtunk adott T_i szöveg $FV_i^j, j = 1, \dots, m$ tulajdonság vektorainak csoportosításához. Minden esetben Koszinusz távolság metrikát használtuk. Ismeretes, hogy a klaszterbe sorolása mérhető mennyiségeknek tulajdonképpen egy kerekítési tevékenység. Ezáltal adott T_i szövegnél az m darab, sorban következő tulajdonság vektor $c_i^{Módszer} < m$ darab klaszterbe sorolódik be, vagyis klaszterezési módszertől függően más-más klaszterszámot kapunk. Az egyes klaszter azonosítókat egy-egy állapotnak tekintjük a Markov lánc számára. Így az $FV_i^j, j = 1, \dots, m$ vektor egy $c_i^{Módszer}$ elemű állapot sorozatot ad, amiben meghatározható az állapotok egymásutánja. Ezt a jelen dolgozat 2. szekciójában tárgyalt $\{Xt\}, t \geq 0$ sorozatnak tekintjük és a keveredési (vegyitési) idő értékét számoljuk. Elsőrendű, illetve másodrendű Markov láncként kezelve az $\{Xt\}, t \geq 0$ sorozatot, a T15 szöveg esetén a három klaszterezési módszer alapján kapott állapot gráf, illetve sajátérték ábrákat az 5., illetve 6. ábra szemlélteti.



5. ábra. A T15 szöveg elsőrendű Markov tranzíciós gráfja és tranzíciós mátrixának sajátértékei: a) DBSCAN algoritmussal; b) Linkage algoritmussal; c) k-Means algoritmussal kapott eredmények.

Megállapítottuk az alábbiakat:

i) A DBScan lényegesen több klaszterbe sorolja be az $FV_i^j, j = 1, \dots, m$ tulajdonság vektorokat, függetlenül a Markov lánc emlékezetétől.

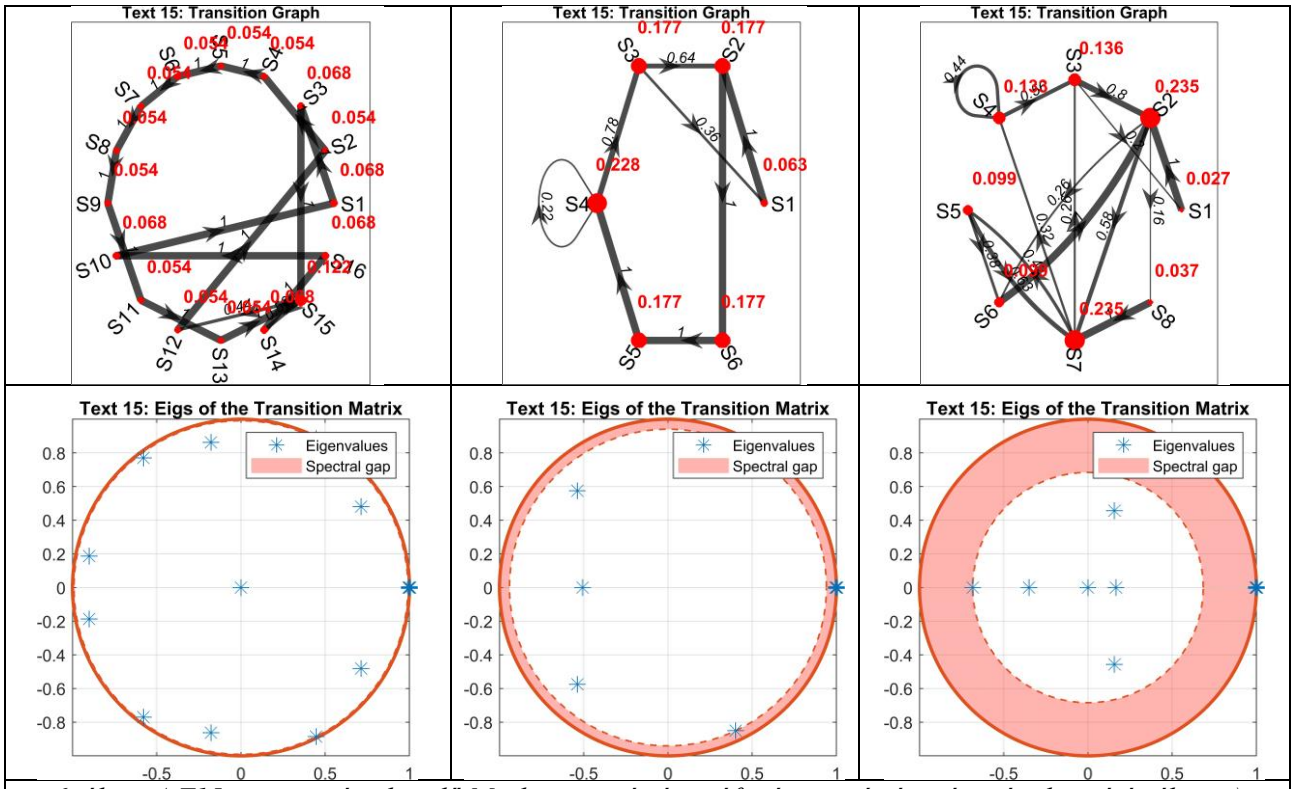
ii) A klaszterek száma a Linkage és a K-Means esetében nem egyértelműen rangsorolható. Ez a sajátértékek darabszámával is egyértelműen igazolható.

iii) Az elsőrendű Markov lánc gyorsabb konvergenciával (kisebb keveredési idővel) rendelkezik, mint a másodrendű Markov lánc.

Egy elsőrendű Markov-láncnak kisebb a keverési ideje, mint egy ugyanazon az állapothalmazon definiált másodrendű Markov-láncnak. Egy elsőrendű lánc csak az aktuális, X_t állapotra emlékszik Egy másodrendű lánc két korábbi állapotra emlékszik (X_{t-1}, X_t), ami gyakorlatilag egy új, n^2 méretű állapotteret hoz létre, ahol n az állapotok darabszáma. A τ_{mix} keverési idő nagyjából az $1 - |\lambda_2|$, spektrális rés reciproka, ahol $|\lambda_2|$ a tranzíciós mátrix második legnagyobb sajátértéke nagyságrendben. Egy másodrendű láncot elsőrendű láncként ábrázolunk n^2 összetett állapotokon. Ez kisebb spektrális réshez vezethet, azaz 1-hez közelebbi sajátértékekhez, mert az (X_{t-1}, X_t) memória lelassítja az információ elfelejtésének sebességét. A lánc bizonyos kétlépéses kontextusokban hosszabb ideig beragadhat. Az állapotgráf gyakran kevésbé összefüggővé és inkább blokkszerkezetűvé válik, ami növeli a korrelációs időt. Az elsőrendű lánc gyorsan

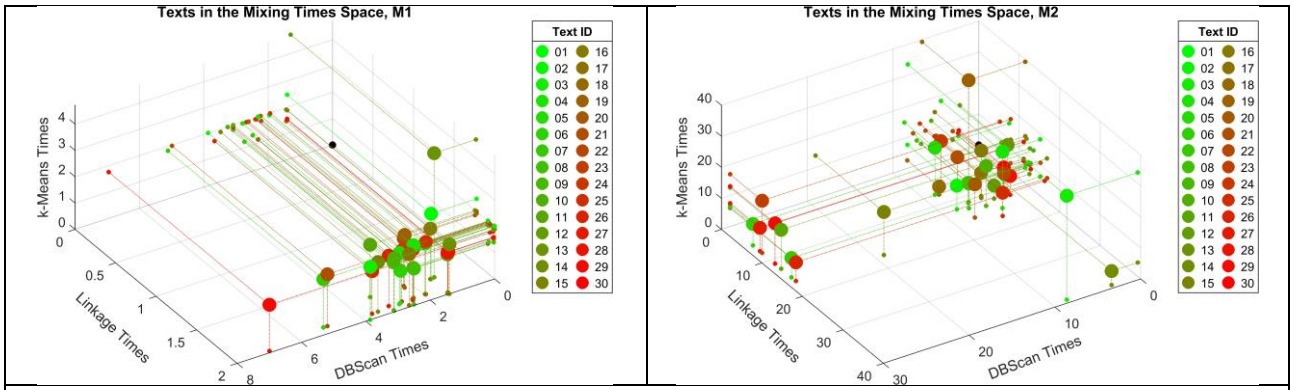
elfelejti a múltat, ezért gyorsabban keveredik. A másodrendű lánc extra memóriát hordoz, ezért lassabban keveredhet.

Trajektória és stílusváltások felismerése terén a jellemzők sorozata gyakran fokozatos stílus- vagy témaváltást mutat. A másodrendű átmenetek a változás irányát (momentumát) is kódolják. Nemcsak azt látjuk, „hol vagyunk” X_t, hanem azt is, „honnan jöttünk” (X_{t-1}, X_t), ami segít felismerni a tendenciákat (pl. növekvő bonyolultság, változó dinamika). A másodrendű modell nemcsak állapotokat, hanem jellemzőátmeneteket is felismer, ami segít azonosítani ismétlődő szerkezeti vagy stílusmintákat.



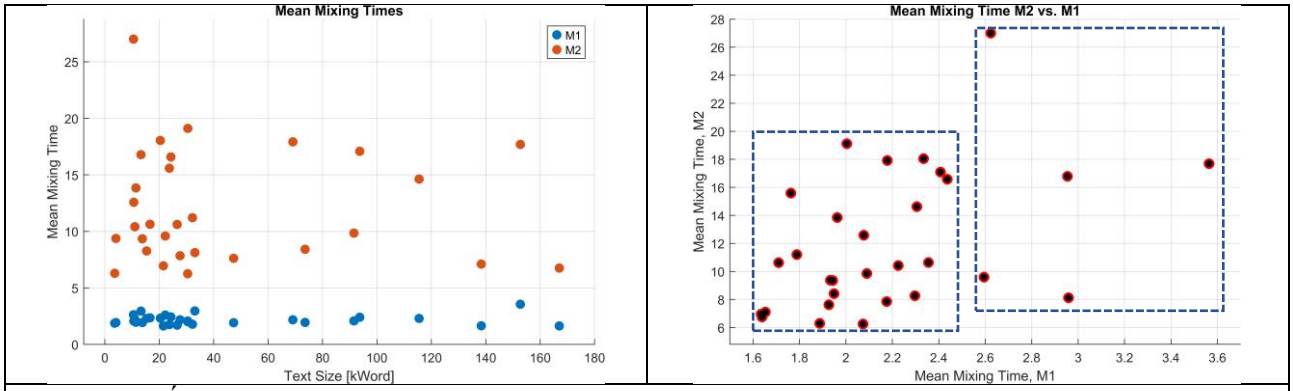
6. ábra. A T15 szöveg másodrendű Markov tranzíciós gráfja és tranzíciós mátrixának sajátértékei: a) DBSCAN algoritmussal; b) Linkage algoritmussal; c) k-Means algoritmussal kapott eredmények.

A keveredési idő értéke függ az alkalmazott klaszterezési módszertől. Ezt szemléltetik a 7. ábra két grafikonja, amelyek az elsőrendű, illetve másodrendű Markov láncok keveredési idejének állapotterében jelenítik meg az N darab szöveget. Elsőfokú Markov láncnál kialakított állapottérben a szövegek egyetlen csoportba sorolhatók (ld. 7. ábra bal oldal). Ez azt igazolja, hogy a CIA témájú szövegek témaköre hasonló megközelítések alapján ismertetik a mondanivalót. A másodrendű Markov láncok esetén készült állapottérben viszont a szövegek két fő csoportra tagolhatók, miközben megjelentek egyedi stílusú írások is. Ez jól mutatja, hogy a másodrendű Markov láncokkal történő elemzés másfajta részletekre is figyel a szövegek dinamikáját illetően (ld. 7. ábra jobb oldal).



7. ábra. Markov lánc vegyítési ideje a DBScan, Linkage és K-Means klaszterezési térben. bal) Elsőrendű Markov lánc; jobb) Másodrendű Markov lánc.

Az átlagos keveredési idő az elsőfokú és másodfokú Markov láncok segítségével a 30 elemzett CIA szövegnél egyértelműsíti a memória hatását. A másodrendű Markov lánc a hosszabb memóriája miatt stabilabban, de lassabban konvergál az állapotváltások egyensúlya felé (ld. 8. ábra bal oldal).



8. ábra. Átlagos vegyítési idők Markov láncoknál (DBScan, Linkage és K-Means klaszterezésekből): bal) Elsőrendű (M1) és másodrendű (M2) Markov lánc a szöveg hossza alapján; jobb) (M1, M2) scatter.

Az átlagos keveredési idők szóródás (scatter) ábrája egy nagyobb és egy kisebb sűrűségű tömörülést mutat (ld. 8. ábra jobb oldal). Ez a jelenség tipikus, ha a szövegek különböző belső rendezettségűek: pl. egyesek inkább véletlen jelsorozat-szerűek, mások szabályos nyelvi szerkezetet mutatnak.

4. ÖSSZEFOGLALÁS

A dolgozat a kvantitatív nyelvészet eszközeivel vizsgálja a szövegek kohézióját és dinamikáját, vagyis azt, hogyan kapcsolódnak össze a szövegrészek és miként bontakozik ki bennük a narratíva. A szöveget egymás utáni szövegegységek (szövegentitások) sorozataként kezeljük, melyeket 17 szófaj kategória arányait tartalmazó tulajdonságvektorokkal írnak le. A vizsgálat 30, a Central Intelligence Agency (CIA) digitális könyvtárából származó, politikai és katonai témájú angol nyelvű szövegen alapul. A szövegentitások és a teljes szöveg globális tulajdonság vektorának hasonlóságát Koszinusz-távolság segítségével számítottuk, amelyből D_i idősort képeztünk, és ennek átlaga adta az adott szöveg átfogó tematikus kohézió (OTC) mérőszámát. Az OTC értékek minden esetben 85% felettiek, és összefüggést mutatnak a szöveghosszal: hosszabb szövegek kohéziója nagyobb. A témaváltások intenzitását a D_i idősor szórása jellemzi, amely szintén a szöveghosszal arányosan nő. A szerzők a szövegdinamika modellezésére első- és Markov-láncokat alkalmaznak, és a szövegek kohéziós szerkezetének statisztikai tulajdonságait a spektrális rés ($1-|\lambda_2|$) segítségével vizsgálják. Eredményeik szerint a szövegek kohéziója és dinamikája kvantitatív módon mérhető, és ezek a jellemzők szorosan összefüggnek a szövegek hosszával és a szerző stílusával. Elsőrendű modell csak az aktuális állapotból jósolja a következőt. A másodrendű modell a két előző állapotot is figyelembe veszi, ezért képes a lokális kontextust jól megragadni, míg az elsőrendű erre nem képes.

KÖSZÖNETNYILVÁNÍTÁS

Ezt a munkát a Debreceni Egyetem QoS-HPC-IoT Laboratórium és a TKP2021-NKTA-34 projekt támogatta. A TKP2021-NKTA-34 számú projekt a Kulturális és Innovációs Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a TKP2021-NKTA pályázati program finanszírozásában valósult meg.

IRODALMI HIVATKOZÁSOK

[1] Tóth, E., Gál, Z. Classification of the Noisy Texts Based on Feature Vectors. In: 15th IEEE International Conference on Cognitive Infocommunications: CogInfoCom 2024: Proceedigns / IEEE, IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC, Piscataway, 77-83, 2024. ISBN: 9798350378245

[2] Gál Z., Tóth E. Deep learning-based analysis of ancient Greek literary texts: A statistical model based on word frequency for the classification of texts. In: *Proc. of the 12th IEEE International Conference on Cognitive*

Infocommunications: CogInfoCom 2021. Ed.: Jan Nikodem, Ryszard Klempous, Piscataway (NJ): IEEE-INST Inc, 2021, 529-535, ISBN: 9781665424950.

[3] Gal Z., Tóth E. Deep Learning-Based Analysis of Ancient Greek Literary Texts in English Version: A Statistical Model Based on Word Frequency and Noise Probability for the Classification of Texts. *Infocommunications Journal*, Joint Special Issue on Cognitive Infocommunications and Cognitive Aspects of Virtual Reality, 2024, 2–11, https://doi.org/10.36244/ICJ.2024.5.1.

[4] Tóth E., Gál Z. Multilabel clustering analysis of the Croatian-English parallel corpus based on Latent Dirichlet Allocation Algorithm. In: *Proc. of the 14th IEEE International Conference on Cognitive Infocommunications: CogInfoCom 2023*, Piscataway (NJ): IEEE-INST Inc, 2023, 25–32, ISBN 97983503256

[5] Tóth E., Gal Z. Optimizing Text Clustering Efficiency through Flexible Latent Dirichlet Allocation Method: Exploring the Impact of Data Features and Threshold Modification. *Infocommunications Journal,* Joint Special Issue on Cognitive Infocommunications and Cognitive Aspects of Virtual Reality, 2024, 58–66, https://doi.org/10.36244/ICJ.2024.5.7.

[6] Phelan, J., Rabinovitz, P. J. Narrative as Rethoric, in Herman, David és mások (eds.): Narrative Theory: Core Concepts and Critical Debates, Columbus, The Ohio State University, 2012, 3-7.

[7] M Jeong, H Kim, SJ Cheon, BJ Choi, NS Kim. Diff-tts: A denoising diffusion model for text-to-speech, arXiv preprint arXiv:2104.01409, 2021, arxiv.org

[8] D. Ghosal, N. Majumder et al. Text-to-Audio Generation using Instruction Guided Latent Diffusion Model, MM '23: Proceedings of the 31st ACM International Conference on Multimedia, pp 3590 – 3598, https://doi.org/10.1145/3581783.3612348.

[9] S. Gu, D. Chen et al. Vector Quantized Diffusion Model for Text-to-Image SynthesisProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10696-10706.

[10] R. Huang, Z. Zhao et al. ProDiff: Progressive Fast Diffusion Model for High-Quality Text-to-Speech MM '22: Proceedings of the 30th ACM International Conference on Multimedia Pages 2595 – 2605, https://doi.org/10.1145/3503161.3547855