Szövegek korrumpálódásának vizsgálata a hálózati adatátvitel során

Text corruption analysis during text transmission on the network

Dr. TÓTH Erzsébet[1], Dr. GÁL Zoltán[2]

[1]Debreceni Egyetem, Informatikai Kar, 4028 Debrecen, Kassai út 26., toth.erzsebet@inf.unideb.hu
[2]Debreceni Egyetem, Informatikai Kar, 4028 Debrecen, Kassai út 26., gal.zoltan@inf.unideb.hu

Abstract

In our paper we propose a method to enhance service levels in unreliable message transmission applications through text corruption detection. Analysis of 20 CIA texts explored that feature vectors arising from part-of-speech tagging show specific patterns independent of text size. Subtexts displayed reduced cohesion with parent texts as their size decreased. An exponential function was fitted to the mean intensity value of sorted token identifiers, presenting a common linguistic structure in English military and political texts. Noise introduction experiments highlighted that higher noise levels decreased the mean coefficient of variation and increased entropy fluctuations of texts' feature vectors, with low noise levels having minimal effect on these metrics. These results reflect that monitoring entropy and coefficient of variation metrics of texts' feature vectors can preserve text cohesion, even under noisy conditions, giving a quantitative framework for understanding text corruption effects and supporting the improvement of reliable communication systems and linguistic models.

Keywords: Central Intelligence Agency (CIA) texts; text classification; part-of-speech-tagging; quantitative linguistics; coefficient of variation; entropy; noisy texts.

Kivonat

Dolgozatunkban egy olyan módszert javaslunk, amely növeli egy nem megbízható üzenetküldő alkalmazás szolgáltatásának minőségét a szövegek korrumpálódásának felderítésével. A Central Intelligence Agency (CIA) szervezet szövegeinek vizsgálata a tokenek mondatrészekbe (Parts of Speech=POS) történő besorolásából származó tulajdonság ("feature") vektorokat tárja fel, amelyek speciális mintázatokat mutatnak a szövegek méretétől függetlenül. A szövegentitások csökkenő kohéziót mutatnak a szülő szöveggel, ahogyan azok mérete csökken. Exponenciális függvényt tudtunk illeszteni a tulajdonság vektorok rendezett token kategóriáinak átlag intenzitás értékére, ami egy gyakori nyelvi struktúrát jelez az angol nyelvű katonai és politikai témájú szövegekben. A zaj bevezetésére irányuló kísérletek magasabb zajszinteknél egyre inkább csökkenő átlag variációs együtthatót eredményeznek a szövegek tulajdonság ("feature") vektoraira vonatkozóan és egyre inkább növekvő átlag entrópia ingadozásokat mutatnak a szövegek tulajdonság vektorainál. Ezzel szemben az alacsony zajszinteknek minimális a hatása a tulajdonság vektorok ezen vizsgált metrikáira. Ezek az eredmények azt tükrözik, hogyha nyomon követjük az entrópia és a variációs együttható metrikákat, akkor a szöveg kohéziója megmarad még zajos feltételek mellett is. Mindez pedig lehetővé teszi, hogy egy megfelelő kvantitatív keretrendszert alakítsunk ki a szöveg korrumpálódás hatásainak megértésére ezzel is támogatva a megbízható hálózati kommunikációs rendszerek és a nyelvi modellek fejlesztését.

Kulcsszavak: Central Intelligence Agency (CIA) szövegek; szövegek osztályozása; mondatrész kategorizálás, kvantitatív nyelvészet; variációs együttható; entrópia; zajos szövegek.

1. BEVEZETÉS

Kutatásunk a kvantitatív nyelvészet területéhez kapcsolódik, amely számos matematikai modellt használ a természetes nyelvű rendszerek elemzésére és a különböző feltárt összefüggéseket számokban, illetve rangsorokban fejezi ki. Ezek az egyetemes, sztochasztikus törvények minden nyelvi szinten és a természetes nyelvekben egyaránt jól alkalmazhatók. A szöveges adatoknál fellépő hibaelemzés az adatintegritást, a hiba felderítését, a hálózati kommunikáció megbízhatóságát, a hálózati mechanizmus hatékonyságát, az adatkonzisztenciát és a felhasználó bizalmát segíti elő. Dolgozatunkban egy olyan vizsgálati módszert javaslunk, ami a hálózaton továbbított szöveges adatok jelentésének torzulását deríti fel a hálózati adatátvitel során bekövetkezett meghibásodáskor, ezzel is elősegítve egy nem megbízható üzenetküldő alkalmazás szolgáltatási minőségének javítását.

Kapcsolódó kutatásként megemlíthetjük az ókori görög irodalmi szövegek klaszterezésének vizsgálatát, amelyet Visszacsatolásos Neurális Hálózat (RNN) segítségével végeztünk el az ókori Alexandriai Könyvtár osztályozási rendszere alapján [1, 2]. Ezen kívül a MARCELL projekt horvát és angol nyelvű párhuzamos jogi korpuszának szövegeit is elemeztük. Megvalósítottuk a korpuszban lévő címkézetlen jogi szövegek hatékony téma besorolásának előrejelzését a Latent Dirichlet Allocation (LDA) algoritmust használva több címkés osztályozásra épülően. Kifejlesztettük az LDA módszer flexibilis változatát a jogi szövegek téma besorolásának javítására, ahol adott küszöbérték megadásával több dobogós téma besorolását tettük lehetővé a jogi szövegek számára [3, 4].

2. ELEMZÉSI MÓDSZERTAN ÉS EREDMÉNYEK

A vizsgált szövegek szerzői események sorozatát írják le adott kronológiai sorrendben, különböző idősíkok között váltva. A szöveg szerzője mindig az egyéni stílusát érvényesítve fogalmazza meg a történetét. Mindez pedig a szöveg kohézióját eredményezi, ami jól felhasználható minden egyes szöveghez tartozó szövegrészlet automatikus jellemzésére, leírására. A dolgozatban adott szöveg egymás utáni mondataiból képezett részletet szövegentitásnak hívjuk, így a szöveg szövegentitások sorozatának tekinthető.

2.1. A feldolgozott szövegek alapvető jellemzői

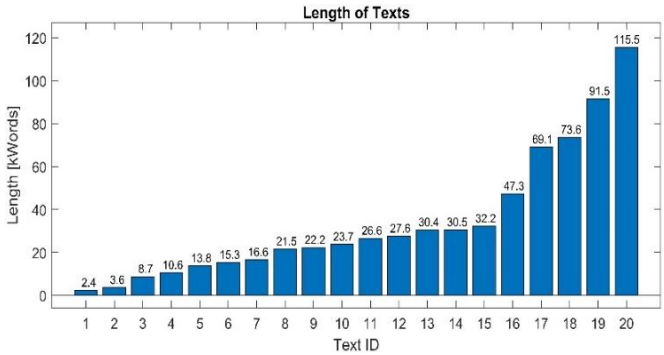
Összesen 20 darab szöveget töltöttünk le a Central Intelligence Agency (CIA) szervezet Digitális Könyvtárából, amelyeket empirikus adatforrásokként kezeltünk. Ezek a szövegek különböző katonai és politikai eseményekről számolnak be, amelyek a világban zajlottak az utóbbi 50 évben. Ebből adódóan a vizsgált szövegeket konzisztensnek és homogénnek tekintettük, mert azok a fent említett témákhoz tartoznak. A szövegeket a szavak számának növekvő sorrendjében adjuk meg (lásd az 1. táblázatot).

A vizsgált CIA szövegek 1. táblázat

	Szerző(k)	Cím
1.	Central Intelligence Agency	Sites to See
2.	Central Intelligence Agency	Memorial Wall Publication
3.	Todd Hazelbarth	The Chinese Media: More Autonomous and Diverse Within — Limits
4.	Central Intelligence Agency	The Work of a Nation: The Center of Intelligence
5.	David W. Waltrop	An Underwater Ice Station Zebra
6.	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
7.	Central Intelligence Agency	A Life in Intelligence - The Richard Helms Collection
8.	Clayton D. Laurie, Andres Vaart	CIA and the Wars in Southeast Asia 1947-75
9.	Central Intelligence Agency	Bosnia, Intelligence, and the Clinton Presidency: The Role of Intelligence and Political Leadership in Ending the Bosnian War

10.	Central Intelligence Agency	Penetrating the Iron Curtain: Resolving the Missile Gap with Technology
11.	Central Intelligence Agency	President Nixon and the Role of Intelligence in the 1973 Arab-Israeli War
12.	Central Intelligence Agency	The Warsaw Pact: Treaty of Friendship, Cooperation and Mutual Assistance
13.	Central Intelligence Agency	Profiles in Leadership: Directors of the Central Intelligence Agency and Its Predecessors 1941-2023
14.	Central Intelligence Agency	Ronald Reagan: Intelligence and the End of the Cold War
15.	Central Intelligence Agency	President Carter and the Role of Intelligence in the Camp David Accords
16.	Andrew Skitt Gilmour	A Middle-East Primed for New Thinking: Insights and Policy Options From the Ancient World
17.	Robert Vickers	The History of CIA's Office of Strategic Research, 1967-81
18.	Thomas L. Ahern, Jr.	"Nothing if Not Eventful": Recollections of a Life's Journey in CIA
19.	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
20.	John L. Helgerson	Getting to Know the President: Fourth Edition: Intelligence Briefings of Presidential Candidates and Presidents-Elect, 1952-2016

A szövegek hossza 2.4 és 115.5 ezer szó közé esik. Az elemzett szövegek 75%-a kevesebb, mint 35 000 szót tartalmaz, ezáltal létrehozva a szövegméretek megközelítőlegesen két lineáris értéktartományát (lásd az 1. ábrát).



1. ábra. Az elemzett szövegek hossza

Mindegyik elemzett szöveget tokenekre (szózsetonokra) konvertáltuk, így lehetővé vált számunkra a szövegek tokenjeinek mondatrész kategória beazonosítása figyelembe véve az adott szöveg kontextusát. A Matlab programozási eszköz függvényei elvégezték számunkra az egyes tokenek megfelelő mondatrész 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

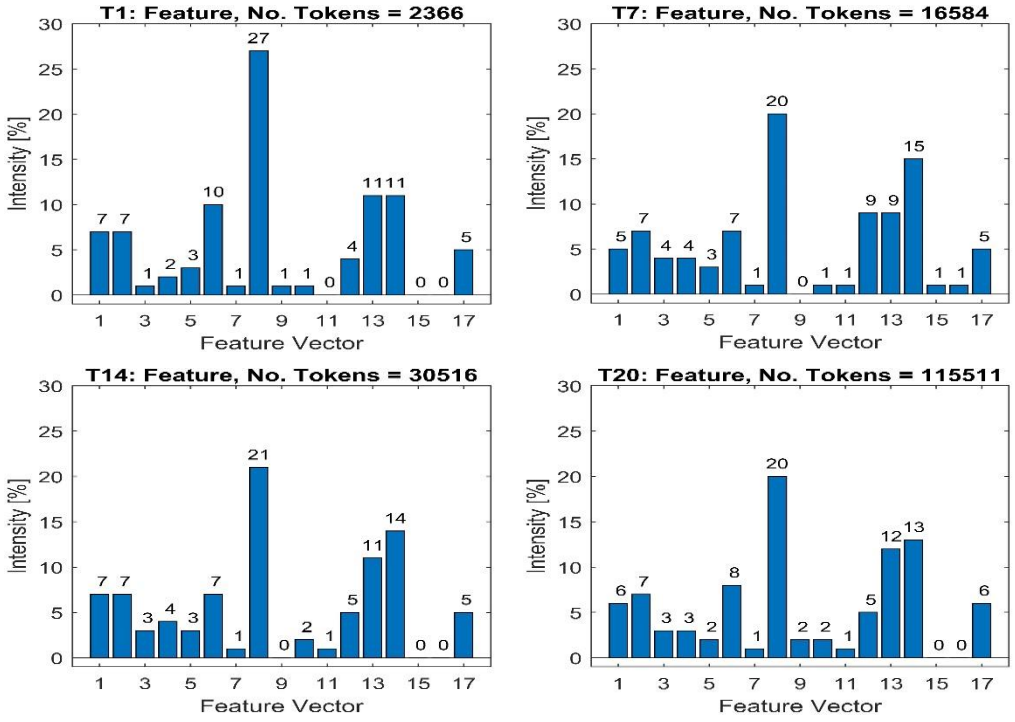
ID	Token kategória
11	viszonyszó
12.	névmás
13	tulajdonnév
14	írásjel
15	alárendelőszó

ID	Token kategória
16	szimbólum
17	ige

A token kategóriák száma összesen 17 volt (lásd a 2. táblázatot). Az "egyéb" token kategória a különböző nemzeti nyelvek lehetséges speciális tulajdonságának kifejezésére szükséges.

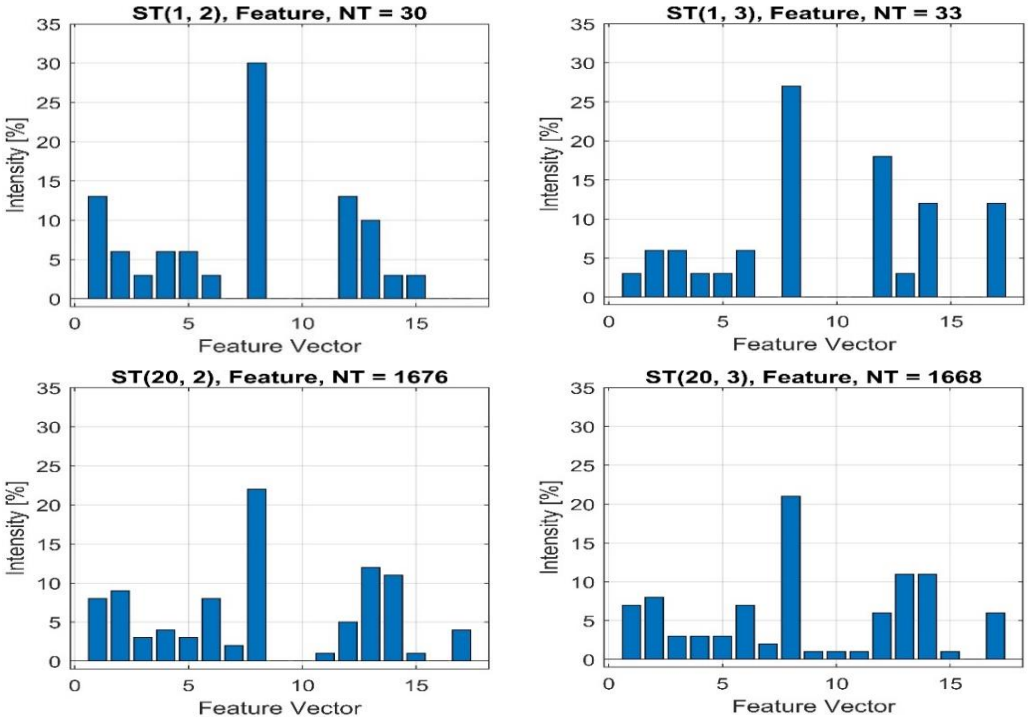
2.2. Szövegek tulajdonság ("feature") vektorainak speciális mintázatai

A tokenek relatív száma a vizsgált szövegekben egy 17 dimenziós tulajdonság ("feature") vektort eredményez, amely a szöveg egészét jellemzi. Megfigyelhető, hogy ezek a vektorok speciális mintázattal rendelkeznek, ami tükrözi az adott szöveg tulajdonságát, jellegét a szövegek méretétől függetlenül (lásd a 2. ábrát).



2. ábra a) b) c) d). Néhány szöveg tulajdonság vektora és a szövegek nincsenek feldarabolva.

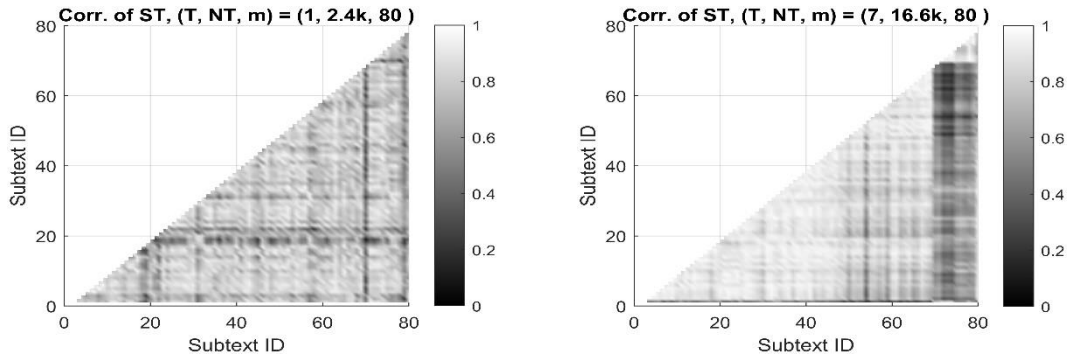
Még az azonos témákba tartozó különféle szövegek is különböző tulajdonság ("feature") vektor mintázattal rendelkeznek a nyelv specifikussága, a szöveg stílusa, illetve szerzője miatt. Néhány kiválasztott szöveg jelentősen eltérő tulajdonság ("feature") vektor mintázata kerül bemutatásra a 2. ábrán. Az intenzitás százalékértékeket lefelé kerekítettük, mert azokat többes címkeként szeretnénk használni egy későbbi kutatásunkban 0 és 100 közötti értéktartományban. A tulajdonság ("feature") vektor példák különböző szövegméretekhez tartoznak, ahol a szöveg ID= 1, 7, 14, 20. Egyértelműen látható a diagramokon, hogy a legnagyobb intenzitás értékkel a főnév token kategória rendelkezik minden egyes esetben. A legritkább token kategóriák pedig a viszonyszó, az alárendelőszó és a szimbólum.

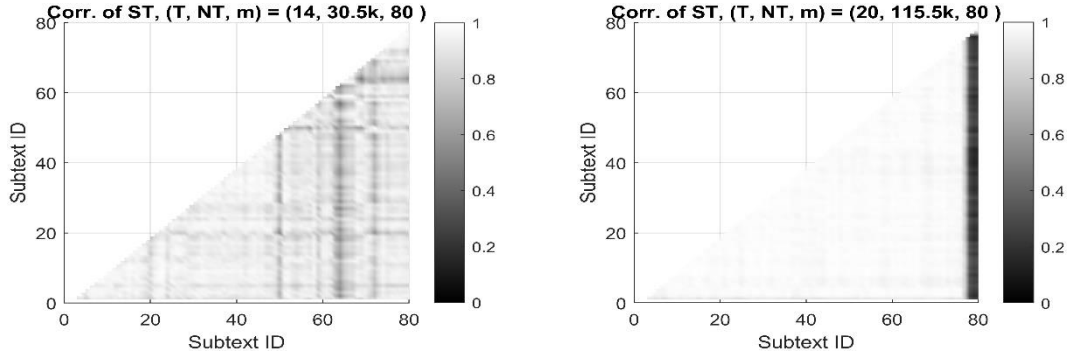


3. ábra a) b) c) d). A különböző szövegekhez tartozó szövegentitások tulajdonság vektorai. Szövegentitások száma/szöveg, m=80

A 20 darab szöveg mindegyik szövegentitására vonatkozóan feldolgoztuk a feature vektorokat. A szövegentitások száma mindegyik szövegben $m \in \mathbb{N}$. Mindegyik szöveg azonos mennyiségű szóra lett felosztva. Mivel a szövegek eltérő hosszúságúak, ezért a különböző szövegekből származó szövegentitások változó hosszúságúak (lásd az 1. ábrát). Megfigyeltük, hogyha minél nagyobb az *m* paraméter, akkor a korreláció annál alacsonyabb a szülő szöveg és annak szövegentitásai között. Ezt a jelenséget szemléltetik a megfelelő diagrampárok (2a, 3a ábrák), (2d, 3d ábrák), stb. Tehát ezt a jelenséget a csökkenő kohézió okozza a szülő szöveg és annak szövegentitásai között. Azaz minél hosszabb a szövegentitás, annál jobban megmarad a kohéziója a szülő szöveggel. A szavak számának csökkenése a szövegentitásban pedig rontja a kontextust. A szövegentitások méretével kapcsolatban egy másik jelenségre figyeltünk fel. Mivel minden könyv az elején tartalmaz egy címoldalt, megjelenéssel kapcsolatos információkat és tartalomjegyzéket, a végén pedig egy tárgymutatót és irodalomjegyzéket, ezért a szöveg felosztásánál az oda eső szövegentitások nagyon eltérnek a fő témától. Ez okozza ezen kis méretű szövegentitások tulajdonság ("feature") vektorainak az eltérését a szülő szöveg tulajdonság vektorától (lásd a 4. ábrát). A világos színű alakzatok 1-hez közel eső korrelációt tükröznek, a sötét színű alakzatok azt jelzik, hogy nincs korreláció a szövegentitások tulajdonság vektorai között.

Újrarendeztük a 20 darab szöveg mindegyik tulajdonság ("feature") vektorának token kategóriáit az intenzitás értékük szerint növekvő sorrendbe, amelyeket rendezett token azonosítóknak, STokenID-nak neveztünk. Felfedeztük, hogy mindegyik szöveg rendezett tulajdonság ("feature") vektora hasonló intenzitás mintázattal rendelkezik a szövegek méretétől függetlenül.



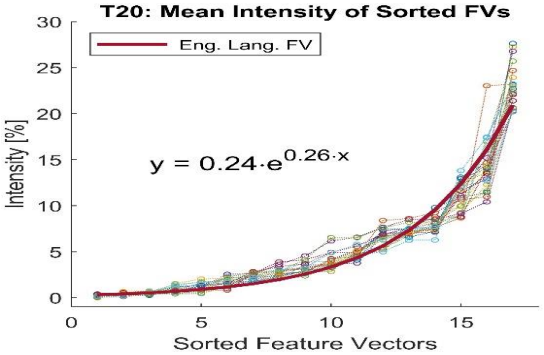


4. ábra a) b) c) d). Néhány szöveghez tartozó szövegentitás tulajdonság vektorai közötti korrelációs együttható: T – szöveg ID; NT – Tokenek száma; m – szövegentitások száma/szöveg

Egy görbét illesztettünk ezen tulajdonság ("feature") vektor mintázatok átlagára és a következő függvényt kaptuk rá:

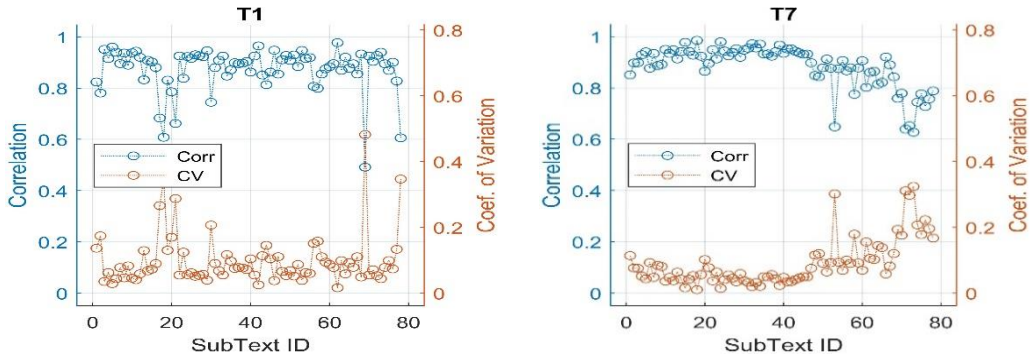
$$Mean\ Sorted\ Token\ Intensity = 0.24 \cdot e^{0.26 \cdot StokenID} \quad (1)$$

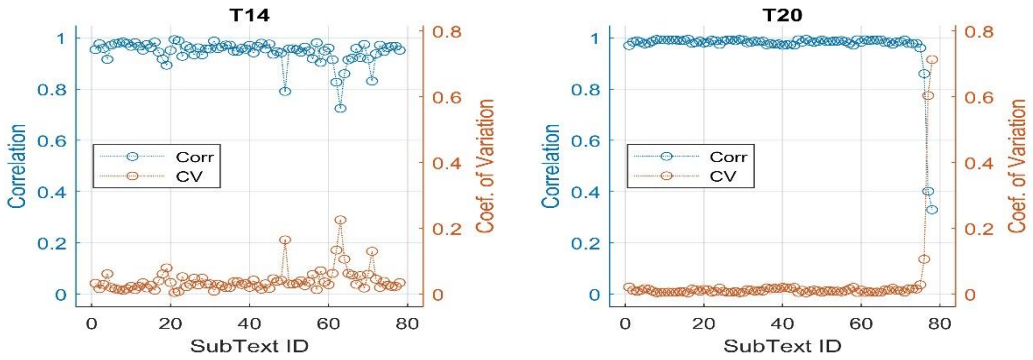
Az 1-es egyenletben exponenciális függvény szerepel, amit MSTI törvénynek nevezünk. Ennek két paramétere gyakran jellemzi az angol katonai és politikai témájú nyelvezetet.



5. ábra. A rendezett tulajdonság vektorok átlaga (angol katonai és politikai témájú nyelvezet)

A variációs együtthatót bármely tulajdonság ("feature") vektor pár hasonlóságának a kvantitatív mérésére használtuk. Az alacsony variációs együttható érték azt jelzi, hogy az adathalmaz tapasztalati szórása kicsi annak átlagához viszonyítva. Ezt azt eredményezi, hogy az adatpontok sűrűn csoportosulnak az átlag körül alacsony változékonyságot tükrözve. Ebből adódóan az adatpontok tehát következetesen hasonlítanak egymásra. Ebben az esetben magas fokú homogenitás létezik az adathalmazban. Alacsony variációs együttható értéknél tehát az adatok megbízhatóak és jobban előrejelezhetőek az alacsony változékonyságuk miatt. Az alacsony variációs együttható érték magas pontosságot is jelez. A különböző szövegekhez tartozó szövegentitások közötti korrelációs és variációs együtthatók szimmetrikusan mozogtak a vízszintes tengely mentén és az $y \approx 0.5$. Ez pedig a két metrika közötti erős függőséget tükrözi (lásd a 6. ábrát). Mindez megerősíti csupán a variációs együttható metrika használatát a kutatásunk folytatásában.

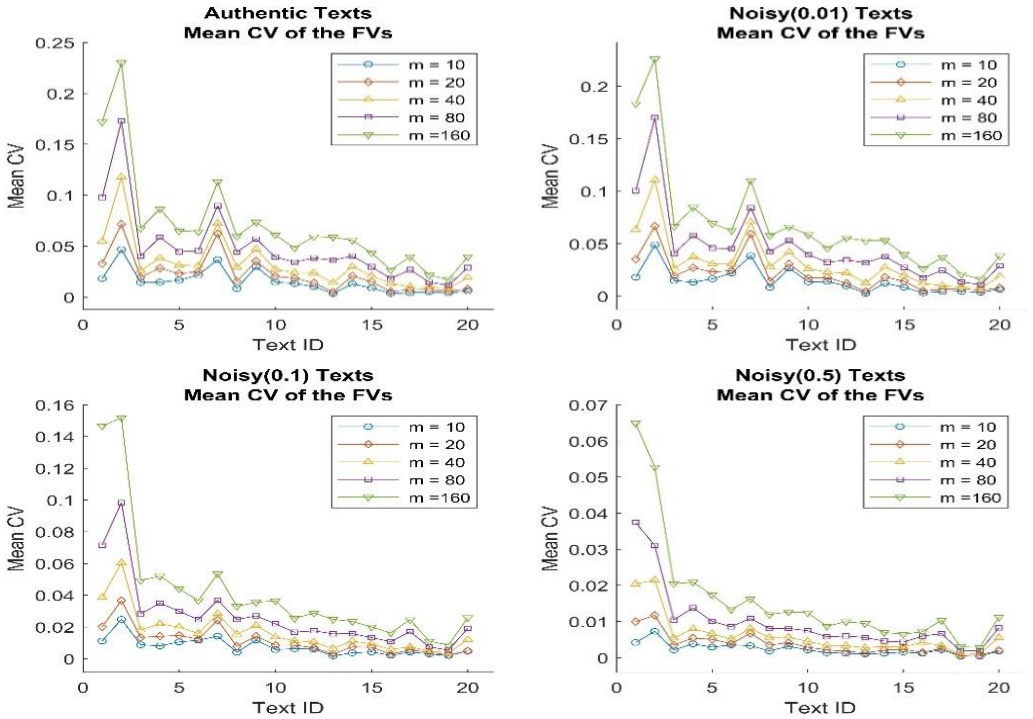




6. ábra a) b) c) d). Néhány szöveg korrelációs és variációs együtthatója

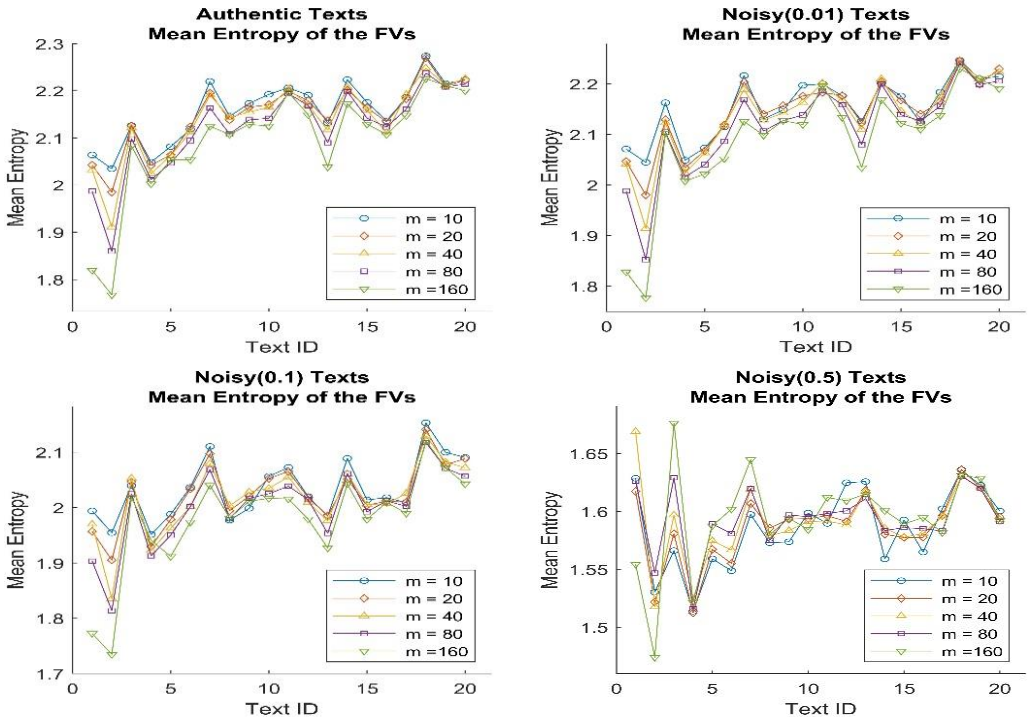
2.3. A random zaj hatása a szöveg tulajdonság vektorok variációs együtthatójára és entrópiájára

Random zajt adtunk hozzá az eredeti CIA szövegekhez, hogy tanulmányozzuk annak hatását az eredeti szövegek jelentésére. Ez a kísérlet a szöveg korrumpálódás hatását vizsgálja a hálózati adatátvitel során. Felfedeztük, hogy mindegyik szöveg tulajdonság ("feature") vektorának variációs együtthatója a szövegentitások számától függ. Azaz minél magasabb az *m* paraméter, annál magasabb lesz a szöveg tulajdonság vektorok átlag variációs együtthatója. A különböző görbék nem metszik egymást (lásd a 7. ábrát).



7. ábra. A szöveg tulajdonság vektorok variációs együtthatójának függése a zajszinttől. a) eredeti szövegek p=0%; b) zajos szövegek p=1%; c) zajos szövegek p=10%; d) zajos szövegek p=50%.

A zajszintet a *p* paraméter befolyásolja, amely a zajos karakterek számának a százalékértékét határozza meg mindegyik szövegben. Minél magasabb a *p* paraméter, annál alacsonyabb a szöveg tulajdonság ("feature") vektorok átlag variációs együtthatója. A nagyon alacsony zajszint (p=1%) nem befolyásolja jelentősen a szöveg tulajdonság ("feature") vektorokat (lásd a 7a és a 7b ábrákat). A 17 valószínűségi változót tekintettük a tulajdonság ("feature") vektorok token intenzitás értékének. Mivel a 17 token kategória egymást kizárja és más egyéb mondatrész kategória nem létezik a második táblázatban a felsoroltakon kívül, ezért mindegyik szöveg tulajdonság ("feature") vektor entrópia metrikáját alkalmaztuk. A szövegfeldolgozásban az entrópia metrika számszerűen fejezi ki egy szöveg rendezetlenségét, ezzel is tükrözve annak információs tartalmát. A magas entrópia érték változatos és kevésbé előrejelezhető szöveget jelent, ezzel szemben az alacsony entrópia érték ismétlődő és előrejelezhető tartalmat jelez.



8. ábra. A szöveg tulajdonság vektorok entrópiájának függése a zajszinttől. a) eredeti szövegek p=0%; b) zajos szövegek p=1%; c) zajos szövegek p=10%; d) zajos szövegek p=50%.

A fenti diagramokon egy általános szabály megfigyelhető melynek értelmében minél nagyobb a szövegek terjedelme, annál nagyobb az entrópia értéke is (lásd a 8. ábrát). Ezt pedig az entrópia (nem) monoton növekedése bizonyítja a szövegméretek függvényében. A görbék sorrendje is megerősíti ezt az állítást a 8a … 8d ábrákon. Ezzel a szabállyal összehangban minél nagyobb a szövegentitások hossza, annál nagyobb az entrópia értéke is. A zajnak sztochasztikus hatása van az entrópia metrikára. Megállapítható, hogy minél nagyobb zajszintnél, egyre nagyobb az entrópia ingadozása is (lásd a 8a … 8d ábrákat). Az alacsony zajszint jelentős mértékben nem befolyásolja a szöveg tulajdonság (“feature”) vektorok entrópiáját.

3. ÖSSZEFOGLALÁS ÉS KÖVETKEZTETÉSEK

A szöveg korrumpálódás vizsgálata elengedhetetlenül fontos az adatintegritás, a hiba detektálás és a hibajavítás biztosítása, valamint a hálózati kommunikáció megbízhatóságának fenntartása miatt. Az egyaránt foglalkozik biztonsággal, protokoll hatékonysággal, adatkonzisztenciával és felhasználói bizalommal kapcsolatos kérdésekkel. A dolgozatban javasolt módszerünk célja, hogy növelje egy nem megbízható üzenetküldő alkalmazás szolgáltatási szintjét a szöveg korrumpálódás detektálására fókuszálva. 20 CIA szöveget vizsgáltunk, a tokenek mondatrész kategóriákba történő besorolásából olyan tulajdonság (“feature”) vektorokat kaptunk, amelyek eltérő speciális mintázattal rendelkeznek a szövegek méretétől függetlenül. A szövegentitások csökkenő kohéziót mutattak a szülő szöveggel, ahogyan azok terjedelme csökkent. Exponenciális függvényt illesztettünk a szövegek rendezett tulajdonság (“feature”) vektorainak átlag intenzitás értékére, ami egy gyakori nyelvi struktúrát jelez az angol nyelvű katonai és politikai témájú szövegekben. A zaj bevezetésére irányuló kísérletek rávilágítottak arra, hogy a magasabb zajszintek csökkentik a szöveg tulajdonság (“feature”) vektorok variációs együtthatóját és növelik azok entrópia ingadozásait. Ezzel szemben az alacsony zajszintek minimális hatást fejtenek ki a szöveg tulajdonság (“feature”) vektorok ezen vizsgált metrikáira. A jövőben még több kutatásra lenne szükség azzal kapcsolatban, hogy a zaj hogyan befolyásolja az entrópia metrikát.

KÖSZÖNETNYILVÁNÍTÁS

Ezt a kutatást a QoS-HPC-IoT Laboratórium és a Debreceni Egyetem TKP2021-NKTA projektje támogatta. A TKP2021-NKTA-34 projektet a Magyarországi Nemzeti Kutatási, Fejlesztési és Innovációs alap támogatta a TKP2021-NKTA finanszírozási formának megfelelően.

IRODALMI HIVATKOZÁSOK

[1] 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

[2] 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

[3] 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

[4] 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

[5] A. Ekbal and S. Bandyopadhyay, Part of Speech Tagging in Bengali Using Support Vector Machine. In: *2008 International Conference on Information Technology*, Bhubaneswar, India, 2008, pp. 106-111, doi: 10.1109/ICIT.2008.12.

[6] Cicero Dos Santos, Bianca Zadrozny, Learning Character-level Representations for Part-of-Speech Tagging, *Proceedings of the 31st International Conference on Machine Learning*, PMLR 32(2):1818-1826, 2014.

[7] Tsuruoka, Y. et al. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: *Bozanis, P., Houstis, E.N. (eds) Advances in Informatics. PCI 2005. Lecture Notes in Computer Science*, Vol. 3746. Springer, Berlin, Heidelberg. https://doi.org/10.1007/1157303636

[8] Peilu Wang, Yao Qian, Frank K. Soong, Lei He, Hai Zhao, Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network, In*: Computer Science, Computation and Language*, (2015) arXiv:1510.06168, https://doi.org/10.48550/arXiv.1510.06168.

[9] Nicolov, Nicolas; Mitkov, Ruslan; Angelova, Galia; Bontcheva, Kalina, Recent Advances in Natural Language Processing III, *John Benjamins Publishing Company - Amsterdam, 2004 - 416 p. - Current Issues in Linguistic Theory* - ISBN: 9789027294685 - Permalink: http://digital.casalini.it/9789027294685 - Casalini id: 5015997.

[10] Chiche, A., Yitagesu, B. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *J Big Data* 9, 10 (2022). https://doi.org/10.1186/s40537-022-00561-y