A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei

Vincze Veronika[1], Kicsi András[1,2], Főző Eszter[3], Vidács László[1,2]

[1]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
[2]Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
Szeged, Dugonics tér 13.
[3]Nemzetbiztonsági Szakszolgálat
Budapest, Törökvész út 32-34.
{vinczev,akicsi,lac}@inf.u-szeged.hu
fozo.eszter@nbsz.gov.hu

Kivonat A modern kommunikációs csatornák használatával a korábbinál jelentősen könnyebbé vált üzenetek anonim módon való közlése közönség, vagy akár kiválasztott emberek számára is. Ez visszaélésekkel is jár, a különböző zsaroló, fenyegető vagy rágalmazó üzenetek forrása is titokban marad. Egyes esetekben a bűnügyi hatóság sem talál erre közvetlen bizonyítékot. A szöveg azonban szükségszerűen tartalmaz bizonyos, a szerzőre jellemző jegyeket. A szöveg stilometriai vizsgálata egy értékes eszköz ilyen nyomozati körülmények között. Ez megköveteli a stílusjegyek azonosítását, amelyet jelenleg a nyelvészeti szakértő a szöveg alapos tanulmányozásával tár fel. Ezt segítő, magyar nyelvű szövegek elemzésére is alkalmas automatizált megoldás nem áll rendelkezésre. Tanulmányunkban azt vizsgáljuk, hogy magyar nyelvű szövegtörzsek automatizáltan kinyert különböző stílusjegyei milyen mértékben utalnak a szerző személyazonosságára, ezzel a későbbiekben hozzájárulva a szakértő munkájához. Az elemzés során számos statisztikai, morfológiai, szintaktikai, szemantikai és pragmatikai jellemző értékeit vetjük össze négy szerző összesen 61 dokumentuma felett, melyek között bűnügyi írások is találhatók. Eredményeink rávilágítanak, hogy a szövegek automatikusan azonosított stílusjegyei alapján lehetséges a szerzők pszichológiai jellemzése, ami a későbbiekben segítheti a bűnügyek felderítését.
Kulcsszavak: szerzőazonosítás, stilometria, kriminalisztika, NLP

1. Bevezetés

Habár a különböző, hang és videó alapú kommunikációs megoldások világunkban igen elterjedtek, az emberek közötti távoli kapcsolattartás többsége napjainkban is szöveges, írott formában történik. Az internet elterjedésével még a korábbinál is könnyebbé, sőt megszokottá vált az írott információ névtelen közlése, amelyen keresztül a személyazonosság akár még az igazságszolgáltatás számára sem, vagy

csak igen nehezen visszakövethető. Az anonim közlésen keresztül ugyan a személyes vélemény szabadabb módon adható át, hozzájárul a szólásszabadsághoz, de rengeteg veszélyt is hordoz. Ilyen esetben a szerzők kevésbé meggondoltan fogalmaznak, sőt akár mások jogait is megsérthetik. Szükség van tehát olyan megoldásokra, amelyekkel ezen visszaélések megnehezíthetők.

Egy írott szöveg szerzőjét felismerni pusztán a szövegre támaszkodva egyáltalán nem triviális feladat. A szöveg óhatatlanul tartalmaz azonban bizonyos nyelvi markereket, amelyek kifejezetten adott szerzőre jellemzőek; a cél ezeknek a nyelvi markereknek a kinyerése, összevetése és az eredmények kiértékelése – automatizált módon.

Két szöveg szerzőjének összerendelése - leegyszerűsítve két szövegtörzs összehasonlítása - továbbra sem egyszerű, a szövegben rengeteg különböző stilometriai jellemző feltárható, amelyek utalhatnak az egyezésre (Michell (2013); Coulthard (1994, 2004)). Ekkora mennyiségű adat pedig szabad szemmel nehezen nyerhető ki és rendszerezhető. Az automatizáció itt utat nyithat egy jelentősen objektívebb látásmód felé, amely kiküszöbölheti az emberi szemlélet és memória hiányosságait. Habár a kriminalisztikai nyelvi elemzés részleges automatizálása több országban már régóta jelen van, az ebben rejlő lehetőségek és lehetséges fejlesztések kiaknázása még messze el van maradva más nyomozati ágak modernizációs törekvéseitől. A német bűnügyi kutatóintézet (Bundeskriminalamt) adatain például történt már hasonló kísérlet (Ishihara (2017)), amelyben a szerző valószínűségi arányokkal (LR, likelihood ratio) végzett kísérletei alapján mutat rá egyes jellemzők információtartalmára a feladat szempontjából. Kísérleteik alapján például a szavankénti átlagos karakterszám, az írásjelek aránya és a szókincs nagysága jó mérőszámoknak bizonyulnak kisebb szövegminták esetén is.

Magyarországon a magyar nyelvű szövegek kriminalisztikai elemzése jelenleg kézzel történik, a szókészlet elemeinek feltárása és összehasonlítása pedig általános konkordancia programokkal (pl. Laurence Anthony szoftverei[1]), ennek automatizációjára tudomásunk szerint korábban nem történt kutatás. A magyar nyelv természetesen továbbra is egyedi problémakört jelent, ami kihat a témában szokásos módszerek felhasználási lehetőségeire. Habár egyes jellemzők használata hasonlóan működik akár a használt nyelvtől függetlenül is, azok információtartalma mégis változhat. Hasonló kiértékelésre tehát szintén szükség van magyar nyelvre is. Ehhez először a nyelvi markerek megállapítására és automatizált kinyerésére kell helyeznünk a hangsúlyt. Az elemzés eredményei pedig a későbbiekben hozzájárulhatnak az intelligens megoldásokhoz is, amelynek során egy gépi megoldás képes lehet objektív módon segíteni nem csak a jellemzők átlátását, hanem magát az egyezésre vonatkozó döntést is.

2. Háttér

Habár külföldön széles körű tudományszervező és aktív kutatási tevékenység folyik a kriminállingvisztika területén (Coulthard és Johnson (2010); McMena-

[1] https://www.laurenceanthony.net/software.html

min (2002); Shuy (2006); Nini (2014); Olsson (2004)), a szövegek kriminalisztikai szempontú gépi elemzésének is nagy bibliográfiája van (pl. Crespo és Frías (2015); Ishihara (2010); Nirkhi és mtsai (2016); Rexha és mtsai (2018); Sousa-Silva (2018); Zhang és mtsai (2014)), sőt nyelveken átívelő megoldásokkal is kísérleteznek (Faqeeh és mtsai (2014); Llorens és Delany (2016)), a magyarországi igazságügyi nyelvész szakértők lényegében a kriminalisztikai szövegnyelvészet máig egyetlen összefoglaló kötete alapján végzik a tevékenységüket (Nagy (1980)). Maga a nyelvész szakértés mibenléte azóta is csak egy viszonylag szűk nyelvész/nyelvészeti érdeklődésű réteget, egy-két cikk, tanulmány erejéig foglalkoztat (pl. Szakácsné Farkas és Jánosné (1988); Pápay (2007); Szegedi (2018); Tolnainé Kabók (2015); Ürmösné Simon (2019); Ránki (2011)). Ennek oka, hogy Magyarországon összesen hat – az Igazságügyi Minisztérium szakértői nyilvántartásába – bejegyzett nyelvész szakértő tevékenykedik, forenzikus intézményben még kevesebb, nem erős a terület lefedettsége. Éppen ezért fontos az NBSZ-nek a szerzőségvizsgálat automatizációjának bevezetése, a megrendelők minél hatékonyabb kiszolgálása érdekében, és ebben megerősítenek minket a külföldi partnerszolgálatok tapasztalatai és a nemzetközi kutatási irányok is.

A kriminalisztikai vagy bűnügyi nyelvészet, még specifikusabban a forensic stylistic/forensic authorship research (Perlman (2018)) alapja az egyéni nyelvhasználat (idiolektus), mellyel nem csak a költők, írók rendelkeznek (Ürmösné Simon (2019)) Az idiolektus az, ahogyan a konkrét személy alkalmazza a nyelvet, mely magában hordozza a nyelvelsajátítás hogyanját és a személy nyelvhez való viszonyulását is. A szocializációs közegekben elsajátítottuk valamilyen mélységben a nyelvet, az adott nyelv (esetünkben a magyar) eszközkészlete valamilyen módon a rendelkezésünkre áll; a nyelv alkalmazása során a fejünkben lévő eszközöket variáljuk, válogatjuk, kombináljuk; a nyelvi kompetenciánkra hatással van szűk környezetünk, társadalmi pozícióink, tanulmányaink, olvasmányélményeink, fogékonyságunk a nyelvre, életkorunk, nemünk stb. (Szilák (1980)). Az egyén nyelvhasználatának bizonyos elemeit adott beszédhelyzethez, adott témához, adott műfajhoz, adott közönséghez stb. igazítja, ugyanakkor bizonyos elemei egyéni, mélyen rögzült, nem feltétlenül tudatos választás eredményei (jellemzően ilyenek a kapcsolóelemek, funkciószavak, mondatalkotási eljárás, állandósult szókapcsolatok használata, helyesírási esetek, ismétlődések, szókészlet egyes elemei stb.). A szakértői munka alapja a fogalmazó általános és különös stílussajátosságainak detektálása.

A nyelvész szakértő munkája során a fenyegetést, becsületsértést, zsarolást, rágalmazást megvalósító névtelen (kérdéses) írásművekből feltárható stílusjegyeket összeveti a gyanúsítottól származó szövegminták stílusjegyeivel, és a hasonlóságok/különbségek mennyiségi és minőségi mutatói alapján következtetéseket von le arra vonatkozóan, hogy a kérdéses írásművek fogalmazója a gyanúsított személy-e vagy sem. Az 1:1 alapú összehasonlítás a legtöbb kriminalisztikai szakértői gyakorlat alapja: adott kérdéses aláírást, okmányt, tárgyat, fotót, beszédet stb. adott mintához; esetünkben a kérdéses írásművek és az összehasonlító szövegminták összevetését jelenti elsősorban morfológiai, szintaktikai, szemantikai és pragmatikai szinten, kvalitatív és kvantitatív módon. A stilometrián felül az

általános korpusznyelvészeti tanulmányokhoz hasonlóan a nyelvész szakértő is alkalmazza azt a technikát, hogy a kérdéses írásművek egyes stílusjegyeit (pl. bizonyos szavak előfordulási gyakoriságát) hasonlítja egy általános korpuszban lévőkhöz (http://mnsz.nytud.hu), hogy megállapítsa, van-e jelentős eltérés a szokásos (átlagos) nyelvhasználattól az adott stílusjegy tekintetében vagy sem (lényegében a Magyar Nemzeti Szövegtárat referenciakorpusznak alkalmazva egy-egy szóelőfordulást, grammatikai jelenséget illetően).

Az online környezetben megvalósított bűncselekmények elkövetői informatikailag legtöbbször lenyomozhatók; ez esetben a nyelvész szakértő bevonása az inkriminált szövegek fogalmazójának megállapítása céljából nem feltétlenül szükséges, hiszen munkája időigényesebb, ezért drágább is. Azoknál a bűnelkövetőknél viszont, akik jól rejtik magukat a digitális térben, a nyelvész szakértők jelenthetik a megoldást, ugyanis képesek az egyre nagyobb számú internetes (verbális és írott formában elkövetett) bűncselekményekhez köthető, online, anonim vagy pszeudonim interakciók nyelvészeti elemzésére. Az internetes szövegek elemzésének igénye pedig új kihívások elé állította a szakértőket. A kommunikációs platformok megváltozásával az inkriminált szövegek is megváltoztak, így a hagyományos kriminalisztikai szövegnyelvészeti felfogás, elemzési módszer sok esetben nem alkalmazható (pl. egy ékezet nélkül írott szövegben az ékezethibák nem látszanak; az IM-szövegekben (Instant Messaging), kommentekben nincs relevanciája a különírás-egybeírásnak, a toldalékok elmaradnak, nincsenek írásjelek, gyakori az ismétlés és a rövidítés stb.), vagyis az idiolektus részének tekinthető nyelvi jellemzők egy része nem nyerhető ki a szövegből. A nyomozó hatóság kérdése a szakértőhöz azonban ezeknél a szövegeknél is változatlan: ki a névtelen szövegek írója?

Az NBSZ szakterület-fejlesztési irányai az alábbiakat tartalmazza:

1. Gépi szövegelemzés: Az első állomás az egy platformon, minél több nyelvi szinten megvalósuló gépi szövegfeldolgozás; automatizáció bevezetése azon szövegek esetében, melyek manuálisan, illetve a már meglévő szoftveres elemzőkkel feldolgozhatók, majd a tapasztalatok fényében megkísérelhető a rövidebb/hiányos szövegekből történő gépi adatkinyerés is.
2. Gépi szöveg-összehasonlítás: A gépi elemző által, mely több szövegszinten is megbízhatóan és hatékonyan dolgozza fel a szöveget, megvalósulhat egy automatikus 1:1 alapú szöveg-összehasonlítás humán, szakértői kontrollal: a gép által feltárt nyelvi jegyek kézzel történő priorálása, vagyis az adott nyelvi jegy megkülönböztető voltának, azonosító erejének meghatározása. A cél, hogy a tesztelést, finomítást követően ez a folyamat is teljesen automatikussá váljon.
3. Gépi szöveg-összehasonlítás meglévő adatbázison: A 2. pont megalapozhatja, előkészítheti az 1:N alapú összehasonlítást, melynek keretében az újonnan keletkezett kérdéses írásműveket össze tudjuk hasonlítani korábbi bűnügyekben keletkezett írásművekkel is, ismétlődéseket, sémákat keresve (pl. fenyegető levelek általános jellegzetességei), vagy akár elkövetői egybeeséseket felismerése is megtörténhet (többszörös bűnelkövetők: ugyanaz a személy

máskor, máshol, másokat, más okból fenyeget/zsarol/rágalmaz). A cél, hogy gépi úton hasonlósági sorrend felállítása történjen meg (score érték alapján).

4. LR-alapú értékelés: A bevezetőben említett, a Bundeskriminalamt által próbált LR-alapú 1:1 szövegösszehasonlítás más kriminalisztikai területen működik (Orbán (2018)), például a nyelvész szakértői területhez legközelebb álló hangtechnikai szakterületen is (pl. Fejes (2018)). A biometrikus azonosító rendszer működéséhez populációs adatbázis kiépítése szükséges, lévén a rendszer a kérdéses hanganyagot hasonlítja a hangmintához és a populációs adatbázisban lévő hanganyagokhoz, majd annak a valószínűségét adja meg, hogy melyiknek nagyobb az esélye: a kérdéses beszélő a mintaadó vagy inkább bárki más. A módszer tehát nagymértékben független a humán szakértőtől: két hipotézis közül a gép határozza meg a valószínűség mértékét, bűnügyi szempontból a gép dönti el, hogy a mintaadó személy bűnös (azonos a kérdéses beszélővel) vagy nem bűnös (a kérdéses beszélő a populációs adatbázishoz jobban illeszkedik). Ennek a módszernek a bevezetése a nyelvész szakértői területre még várat magára; a folyamatnál kulcsfontosságú a populációs adatbázis összetétele és a szövegekből az adatkinyerés hatékonysága, pontossága.

Jelen kutatásunk első állomása a gépi elemző és összehasonlító kidolgozásáról szól, a minél szélesebb körű és pontosabb adatkinyerés megvalósításáról, egyre kevesebb humán részvétellel, valamint arról, hogy a kinyert adatokat statisztikai szinten összehasonlíthatóvá tegyük a – pusztán statisztikai adatok alapján számított – gépi értékítélet megfogalmazásához a fogalmazó-szerzők azonosságának/különbözőségének valószínűségét illetően.

3. Módszerek

Jelen tanulmányban azt vizsgáljuk, hogy milyen típusú nyelvi markerek játszanak fő szerepet a szövegek szerzőinek azonosításában. Ehhez egy kisebb korpuszt készítettünk, mely négy különböző szerzőtől tartalmaz különböző írásműveket. Ezek között vannak "hétköznapi" jellegű dokumentumok is, de vannak különféle bűnügyekhez kapcsolódó írások is (pl. zsaroló vagy fenyegető levelek.) A korpusz alapvető adatai az 1. táblázatban láthatók.

1. táblázat. A kísérletek során felhasznált korpusz adatai.

Szerző	Tokenszám	Mondatszám	Dokumentumszám
A	7127	555	10
B	3205	170	6
C	6323	461	13
D	15791	912	32
Összesen	32446	2098	61

E tanulmányban arra keressük a választ, hogy akár egy kis elemszámú mintán is kimutathatók-e olyan (szignifikáns) különbségek, amelyek egyértelműen jellemzik egy szerző idiolektusát, ami a szerzőség megállapításában fontos szerepet játszhat. A szövegeket a magyarlanc nyelvi elemzővel (Zsibrita és mtsai (2013)) elemeztük morfológiai és szintaktikai szinten, valamint különféle szótárak alapján a szókincsüket is részletes elemzésnek vetettük alá. A kapott elemzésből automatikusan nyertük ki az alábbi jellemzőket.

Statisztikai jellemzők: tokenek száma, mondatok száma, lemmák száma és aránya, mondatok átlagos hossza, csupa nagybetűből álló szavak száma és aránya, nagy kezdőbetűs szavak száma és aránya, kijelentő mondatok száma és aránya, felszólító/felkiáltó/óhajtó mondatok száma aránya, kérdő mondatok száma és aránya, a szöveg telítettsége (lemmaszám / tagmondatok száma)

Morfológiai jellemzők: Főnevek, igék, melléknevek, ismeretlen szavak, határozószavak, tulajdonnevek, számnevek, névmások, vonatkozó és mutató névmások, névutók és központozás száma és aránya, múlt és jelen idejű, feltételes és felszólító módú, gyakorító, műveltető és ható, adott számú és személyű igék száma és aránya, középfokú és felsőfokú melléknevek száma és aránya, többes számú főnevek száma és aránya, különleges képzővel/végződéssel rendelkező szavak száma és aránya

Szintaktikai jellemzők: Alanyok, tárgyak, jelzők, határozók, alárendelések, mellérendelések száma és aránya, tagmondatok száma és aránya, egyszerű mondatok száma és aránya, összetett mondatok száma és aránya, egy, két, három vagy négy tagmondatból álló mondatok száma és aránya

Szemantikai jellemzők: Pozitív és negatív töltetű szavak száma és aránya (Szabó (2015) alapján), negatív emotív szavak száma és aránya, tagadószavak, funkciószavak és tartalmas szavak száma és aránya, trágár és rasszista szavak száma és aránya, speciális stílusértékű szavak száma és aránya, megszólítások, elköszönő formulák és utóiratok száma és aránya, bizonytalanságra (Vincze (2014)) és érzelmekre (Szabó és Vincze (2016)) utaló szavak száma és aránya

Pragmatikai jellemzők: beszédaktusok száma és aránya, idézetek és gondolatjelek száma és aránya, kifelé és befelé forduló igék száma és aránya, meggyőzést jelentő igék száma és aránya, diskurzusjelölők száma és aránya

Ezeken felül még szókincselemzést is végeztünk a leggyakoribb szavak jelentésmezőinek összevetésével.

4. Eredmények

Az alábbiakban csak a legfontosabb eredményekre koncentrálva mutatjuk be vizsgálataink eredményét. Mivel a szerzőktől eltérő nagyságú minta állt rendelkezésünkre, különös tekintettel arra, hogy szövegeink közel fele a D szerzőtől származik, elsősorban az egyes nyelvi jelenségek arányaira összpontosítunk, nem a darabszám szerinti előfordulásra. Ugyanakkor megemlítjük, hogy elvégzett statisztikai szignifikanciatesztjeink szerint a négy szerző nyelvhasználata szignifikáns eltérést mutat egymástól (ANOVA, F=13,2948, p=2,23017E-08). Az

alábbiakban felsorolt jellemzők kinyerése mind automatizált módon, szoftveresen történt, ezen jellemzők kinyerése tehát jól automatizálható.

4.1. Statisztikai jellemzők

A 2. táblázat mutatja a legfontosabb statisztikai jellemzők számszerűsített eredményeit. Érdekes megfigyelni, hogy itt is már nagy különbségek mutatkoznak a négy szerző között: B szerző kiemelkedően hosszú mondatokat ír, ami megmutatkozik az átlagos mondathosszban és a mondatonként tagmondatok számában is. A szerzőnél különösen nagy a felszólító mondatok száma, míg C és D szerző viszonylag gazdagabb szókinccsel rendelkezik a másik két szerzőnél.

2. táblázat. A jelentősebb statisztikai jellemzők eredményei.

Jellemző	A	B	C	D
Lemmák aránya	0,48	0,44	0,57	0,58
Mondathossz	11,49	19,33	12,79	16,89
Kijelentő mondatok aránya	0,56	0,35	0,41	0,69
Felszólító mondatok aránya	0,30	0,11	0,09	0,09
Kérdések aránya	0,07	0,10	0,05	0,01
Telítettség	2,58	4,06	3,47	3,29

4.2. Morfológiai jellemzők

A 3. táblázat mutatja a legfontosabb morfológiai jellemzők számszerűsített eredményeit. Az egyes szófajok használati aránya is változik szerzőnként, illetve szerzőpáronként: úgy fest, hogy B és C, valamint A és D szerzők szófajhasználata viszonylag közel áll egymáshoz páronként. A múlt és jelen idő használata terén ugyanakkor B szerző tér el szignifikánsan a többiektől: nála jóval gyakoribb a múlt idő, azaz talán gyakrabban emlegeti a múltbeli cselekvéseket, míg a többi szerző inkább a jelenre/jövőre fókuszál. Noha A szerzőnél figyelhettük meg a felszólító/felkiáltó mondatok nagyobb gyakoriságát, a morfológiai elemzések szerint D szerző használ nagyobb arányban felszólító módú igéket. E két tényező összevetése mindenképpen további vizsgálatokat indokol. Az igei személyragozást vizsgálva is nagy különbségeket láthatunk a szerzők között. Míg B szerző szinte csak önmagáról, illetve harmadik személyekről beszél (a második személyű igealakok gyakorlatilag teljesen hiányoznak a szövegeiből), addig A és D szerző gyakran használ E/2. formulákat, azaz direktben megszólítja a szövegek címzettjét. Nem zárhatjuk ki természetesen, hogy B szerző magázva szólítja meg a címzetteket, de ennek igazolása további vizsgálatokat kíván.

4.3. Szintaktikai jellemzők

A 4. táblázat mutatja a legfontosabb szintaktikai jellemzők számszerűsített eredményeit. Itt nagyobb különbségeket az alá- és mellérendelések, valamint a hatá-

3. táblázat. A jelentősebb morfológiai jellemzők eredményei.

Jellemző	A	B	C	D
Főnevek aránya	0,19	0,28	0,27	0,19
Igék aránya	0,15	0,07	0,09	0,14
Melléknevek aránya	0,07	0,11	0,10	0,07
Határozószavak aránya	0,13	0,05	0,06	0,11
Tulajdonnevek aránya	0,00	0,05	0,03	0,02
Kötőszavak aránya	0,08	0,05	0,04	0,07
Írásjelek aránya	0,19	0,25	0,27	0,19
Névmások aránya	0,08	0,04	0,04	0,09
Múlt idő aránya	0,24	0,40	0,23	0,22
Jelen idő aránya	0,67	0,49	0,66	0,71
Feltételes mód aránya	0,08	0,07	0,04	0,04
Felszólító mód aránya	0,07	0,04	0,08	0,12
E/1. aránya	0,28	0,33	0,25	0,26
E/2. aránya	0,14	0,00	0,05	0,18
E/3. aránya	0,40	0,39	0,44	0,35
T/1. aránya	0,05	0,00	0,09	0,03
T/2. aránya	0,00	0,00	0,00	0,00
T/3. aránya	0,05	0,16	0,07	0,10

rozók használata terén láthatunk. B és C szerző a többiekhez képest gyakrabban használ mellérendelést, míg A szerző inkább az alárendelést és a határozók használatát részesíti előnyben. Az egyes mondatok összetettségét leíró mutatók alapján elmondhatjuk, hogy A és D szerző inkább összetett mondatokat, míg B és C szerző inkább egyszerű mondatokat használ. Ugyanakkor B szerzőnél kiemelkedik a négy tagmondatot tartalmazó mondatok aránya, ami arra utal, hogy noha az egyszerűbb mondatok gyakoribbak nála, ha mégis több tagmondatot foglal egybe, akkor az az átlagosnál hosszabb mondatot eredményez.

4. táblázat. A jelentősebb szintaktikai jellemzők eredményei.

Jellemző	A	B	C	D
Alárendelés aránya	0,05	0,02	0,03	0,04
Határozók aránya	0,05	0,01	0,02	0,03
Mellérendelés aránya	0,05	0,10	0,09	0,07
Tagmondatok száma	2,15	2,10	2,08	3,02
Egyszerű mondatok aránya	0,47	0,61	0,55	0,44
Összetett mondatok aránya	0,53	0,39	0,45	0,56
Egy tagmondatos mondatok aránya	0,47	0,61	0,55	0,44
Két tagmondatos mondatok aránya	0,15	0,08	0,10	0,13
Három tagmondatos mondatok aránya	0,24	0,07	0,17	0,13
Négy tagmondatos mondatok aránya	0,09	0,15	0,08	0,12

4.4. Szemantikai jellemzők

Az 5. táblázat mutatja a legfontosabb szemantikai jellemzők számszerűsített eredményeit. Várakozásainkkal némileg ellentétben, alig találhatunk különbséget a szerzők között e téren, szinte csak minimális eltéréseket mutat egy-egy szemantikai jegy. Noha a kriminalisztikai jelleg miatt például várhatnánk a negatív érzelmekre (pl. düh, frusztráció, fenyegetés) utaló szavak felbukkanását a szövegekben, valójában ezek csak minimálisan fordulnak elő a szövegekben. Természetesen előfordulhat az is, hogy a rendelkezésre álló szentiment- és emóciószótárak nem ilyen típusú szövegekre lettek felkészítve, emiatt nem tudjuk azonosítani a szövegekben rejlő emóciókat, vagy pedig a minta kis mérete nem teszi lehetővé ilyen jellegű különbségek kimutatását. A szövegek részletesebb szemantikai vizsgálata tehát mindenképpen indokolt a jövőben.

5. táblázat. A jelentősebb szemantikai jellemzők eredményei.

Jellemző	A	B	C	D
Pozitív szavak aránya	0,03	0,01	0,03	0,03
Negatív szavak aránya	0,02	0,02	0,02	0,03
Tagadás aránya	0,03	0,01	0,01	0,02
Funkciószavak aránya	0,46	0,44	0,46	0,47
Tartalmas szavak aránya	0,54	0,56	0,54	0,53
Feltételes szavak aránya	0,01	0,00	0,00	0,01
Weasel szavak aránya	0,02	0,01	0,01	0,02
Peacock szavak aránya	0,01	0,00	0,00	0,00
Hedge szavak aránya	0,01	0,00	0,01	0,01
Doxasztikus szavak aránya	0,01	0,00	0,01	0,01
Szorongás szavainak aránya	0,01	0,00	0,00	0,00
Öröm szavainak aránya	0,01	0,00	0,01	0,01

4.5. Pragmatikai jellemzők

A 6. táblázat mutatja a legfontosabb pragmatikai jellemzők számszerűsített eredményeit. Míg a kifelé forduló igék aránya nem mutat lényegesebb eltérést, addig B szerző feltűnően kevés befelé forduló igét használ, ő tehát inkább a külvilágra, semmint önmagára fókuszál írásaiban. Ugyanakkor ő használja arányaiban a legtöbb meggyőzésre utaló szót, ami ugyanakkor azt erősíti, hogy saját véleménye mellett érvel, azt ebben a formában kifejezésre juttatva. Érdekes még megfigyelni a diskurzusjelölők eltérő gyakoriságát: A és D szerzők használják gyakrabban, míg B és C szerzőknél némileg háttérbe szorul a szerepük.

4.6. Szókészleti jellemzők

Az alábbi ábrákon bemutatjuk az egyes szerzők által használt szavak leggyakoribb elemeit, szófelhő formájában.

6. táblázat. A jelentősebb pragmatikai jellemzők eredményei.

Jellemző	A	B	C	D
Kifelé forduló igék aránya	0,04	0,07	0,05	0,04
Befelé forduló igék aránya	0,15	0,06	0,12	0,16
Meggyőzést jelentő igék aránya	0,03	0,08	0,06	0,03
Diskurzusjelölők aránya	0,10	0,04	0,04	0,11

1. ábra: Bal oldal: A szerző szókincsének leggyakoribb elemei, Jobb oldal: B szerző szókincsének leggyakoribb elemei

Az A szerző (1. ábra bal oldala) szókincse alapján elsődleges családcentrikus személynek tűnik, szövegeiben legalábbis túlnyomó többségben fordulnak elő a család témakörhöz kapcsolódó szavak (*család, feleség, férj, házasság...*). Emellett a hétköznapi élet szavai is gyakoriak (*lakás, szoba, telefon, ajtó* stb.).

B szerző (1. ábra jobb oldala) szókincsének leggyakoribb elemeit két fő csoportra oszthatjuk. Az egyik sajátos csoportba sorolhatók a vadászattal kapcsolatos kifejezések (*vadászat, vadászjegy, dám*), míg a másik csoportot a hivatali nyelv jelenti (*határozat, földművelésügyi, hivatal...*). Érdemes megfigyelni, hogy ugyanakkor megjelennek a bűnügyi kifejezések is, mint pl. bűncselekmény, meghamisítás, feljelentés, törvénysértő stb.

A C szerző (2. ábra bal oldala) szókincsében megint csak dominálnak a bűnüggyel kapcsolatos szavak: *feljelentés, csalás, BTK, bűncselekmény*, ugyanakkor a munkaügy-hivatali szókincs is erősen jelen van: *igazgató, APEH, gazdasági, munkatárs, revizor* stb.

D szerző (2. ábra jobb oldala) szókincsének leggyakoribb elemei szintén rendőrségi ügyekkel kapcsolatosak: *nyomozás, szakértői, rendőrség, vallomás, ügy, ujjlenyomat, fenyegető, iratismertetés* stb. Ez talán arra utal, hogy számára sem ismeretlen egy rendőrségi ügy lezajlásának miként je.

5. Az eredmények megvitatása

Az előző fejezetben bemutatottak szerint létezik néhány nyelvi jellemző, amely hatékonyan működhet a szövegek szerzőinek profilozásában, illetve azonosítá-

2. ábra: Bal oldal: C szerző szókincsének leggyakoribb elemei, Jobb oldal: D szerző szókincsének leggyakoribb elemei

hoz is hatékonyan hozzájárulhat. A jellemzők alapján az alábbiakat állapíthatjuk meg az egyes szerzőkre nézve. A szerző nyelvhasználatában a hétköznapi élet szókincse dominál, valószínűleg kevésbé iskolázott a többi szerzőnél. Különösen nagy nála a felszólító vagy felkiáltó mondatok száma, gyakran használ E/2. ragozást, azaz direktben megszólítja az olvasót. Igeidőket tekintve inkább jelen vagy jövőbeli eseményekre fókuszál, kevésbé említi a múltat. Diskurzusjelölők is gyakran fordulnak elő a szövegeiben, gyakran használ összetett (alárendelő) mondatokat, így írásai inkább az élőbeszédre emlékeztetnek, semmint hivatalos szövegekre. B szerző mondatai kiemelkedően hosszúak, gyakran használ mellérendelést, ugyanakkor inkább az egyszerűbb mondatstruktúrát részesíti előnyben. A többi szerzőhöz képest ő többször említ múltbeli eseményeket, szinte csak E/1. vagy E/3. ragozást használ, azaz nem közvetlenül egy másik (tegezett) személynek intézi a mondandóját. Tetten érhető nála a külvilágra való fókuszálás, ugyanakkor azon tendencia is, hogy szeretné meggyőzni a szövegek címzettjét a saját véleményéről, feltehetően érvelő jelleggel próbálja bizonyítani igazát. A szövegek tematikája esetében nagyon speciális, a vadászat szókincsének erős jelenléte mindenféleképpen különleges támpontot adhat a szerző kilétére irányuló nyomozás során. Ugyanakkor az is kiderül, a hivatalok és rendőrségi eljárások világa sem ismeretlen számára. C szerző - B-hez hasonlóan - szintén inkább egyszerűbb mondatokat, de több mellérendelést használ, szókincse változatosnak mondható. Diskurzusjelölőket ő is ritkábban használ, nyelvezete hivatalosabbnak tűnik, mint pl. A szerzőé. Szókincsét tekintve nála is a munka és a bűnügyek világa dominál. Nyelvezete inkább B-éhez áll közelebb. D szerző szintén gazdag szókinccsel rendelkezik. Inkább összetett mondatokban kommunikál, relatíve sok diskurzusjelölőt alkalmaz. Ő is gyakran használ E/2. alakokat, azaz megszólítja a címzettet, a jelenre vagy jövőre fókuszál. E sajátságok inkább az élőbeszédi kommunikációra emlékeztetnek, így D nyelvezete (és profilja) talán A-éhoz áll legközelebb. Szókincsének alapján ő is tisztában van a rendőrségi eljárásokkal.

Természetesen a fenti következtetések csak segítik a nyomozók munkáját, önmagukban nem rendelkeznek bizonyító erővel az eljárásokban. Munkacsopor-

tunk jelenleg azon dolgozik, hogy a fenti jellemzőket automatikusan kinyerve a szerzőktől rendelkezésre álló szövegekből, azokat számszerűen összehasonlítva és azt egy nyelvész szakértő elé tárva, a gépi úton kinyert adatok automatikus összevetésével segítsük a szerzőazonosítás folyamatát.

6. Összegzés

E munka célja annak megmutatása volt, hogy egy kis mintán is lehetséges automatikus nyelvi elemzési eszközökkel olyan összefüggéseket felállítani, amelyek segítik a kriminalisztikai célú szerzőazonosítást. Felhívtuk a figyelmet több olyan nyelvi jellemzőre, melyek már akár kis szövegminta alapján is árulkodónak bizonyulhatnak a szerző stílusára nézve, illetve segíthetik a más szerzőktől való elkülönítést. További célunk, hogy a szerzőazonosítás folyamatát minél inkább automatizáljuk, valamint meghatározzuk az elemzett adatok gépi összehasonlítása alapján a fogalmazó-szerzők azonosságának/különbözőségének valószínűségét, hozzájárulva ehhez a nyomozói munka hatékonyságához és sikerességéhez.

Köszönetnyilvánítás

A publikációban szereplő kutatást (amelyet a Nemzetbiztonsági Szakszolgálat és a Szegedi Tudományegyetem valósított meg) az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

Hivatkozások

Coulthard, M.: On the use of corpora in the analysis of forensic texts. In: Forensic Linguistics: The International Journal of Speech, Language and the Law. vol. 1, p. 27–43 (1994)

Coulthard, M.: Author identification, idiolect, and linguistic uniqueness. In: Applied Linguistics. vol. 25, p. 431–447 (2004)

Coulthard, M., Johnson, A.: The routledge handbook of forensic linguistics (2010)

Crespo, M., Frías, A.: Stylistic authorship comparison and attribution of spanish news forum messages based on the treetagger pos tagger. In: Procedia - Social and Behavioral Sciences. p. 198–204. No. 212 (2015)

Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., Quwaider, M.: Cross-lingual short-text document classification for facebook comments. In: International Conference on Future Internet of Things and Cloud. p. 67–98 (2014)

Fejes, A.: Beszéd alapján történő személyazonosítás új kihívásai a kriminalisztikában. In: Magyar Rendészet. p. 117–126. No. 2 (2018)

Ishihara, S.: E-mail authorship verification for forensic investigation. In: Proceedings of the 2010 ACM Symposium on Applied Computing (SAC). vol. 24. Sierre, Switzerland (2010)

Ishihara, S.: Strength of forensic text comparison evidence from stylometric features: A multivariate likelihood ratio-based analysis. In: International Journal of Speech Language and the Law. vol. 24, p. 67–98 (2017)

Llorens, M., Delany, S.: Deep level lexical features for cross-lingual authorship attribution. In: Proceedings of the First Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine 2016). Padova, Italy (03 2016)

McMenamin, G.: Forensic linguistics. advances is forensic stylistics. CRC Press (2002)

Michell, C.S.: Investigating the use of forensic stylistic and stylometric techniques in the analyses of authorship on a publicly accessible social networking site (facebook). In: Dissertation. University of South Africa (2013)

Nagy, F.: Kriminalisztikai szövegnyelvészet. Akadémiai Kiadó, Budapest (1980)

Nini, A.: Authorship profiling in a forensic context. In: PHD thesis (2014)

Nirkhi, S., Dharaskar, R., Thakare, V.: Authorship verification of online messages for forensic investigation. In: Procedia Computer Science. vol. 78, p. 640–645 (2016)

Olsson, J.: Forensic linguistics: An introduction to language, crime, and the law. Bloomsbury Publishing, New York (2004)

Orbán, J.: Bayes-hálók a bűnügyi tudományokban. In: PhD értekezés. Pécs (2018)

Perlman, A.: What is forensic stylistics? (2018), https://www.language-expert.net/category-stylistic-analysis

Pápay, K.: Valószínűségi skálák az igazságügyi nyelvészetben. In: I. Alkalmazott Nyelvészeti Doktorandusz Konferencia kötet. p. 102–113. MTA Nyelvtudományi Intézet, Budapest (2007)

Rexha, A., Kröll, M., Ziak, H., Kern, R.: Authorship identification of documents with high content similarity. In: Scientometrics. vol. 115, p. 223–237 (2018)

Ránki, S.: A kriminalisztikai szövegnyelvészet hazai kutatástörténete 1960-tól 1990-ig. In: E-nyelvmagazin (2011), https://e-nyelvmagazin.hu/2011/08/31/a-kriminalisztikai-szovegnyelveszet-hazai-kutatastortenete-1960-tol-1990-ig/

Shuy, R.: Linguistics in the courtroom: A practical guide. Oxford University Press, New York (2006)

Ürmösné Simon, G.: Miben segítik a nyelvi ujjnyomok a nyomozást? In: Magyar Rendészet. p. 65–75. No. 1 (2019)

Sousa-Silva, R.: Computational forensic linguistics: An overview of computational applications in forensic contexts 5, 118–143 (12 2018)

Szabó, M.K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. nyelv, kultúra, társadalom. In: Segédkönyvek a nyelvészet tanulmányozásához. p. 278–285 (2015)

Szabó, M.K., Vincze, V.: Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: Távlatok a mai magyar alkalmazott nyelvészetben. p. 282–292. Tinta pp., Budapest (2016)

Szakácsné Farkas, J., Jánosné, V.: A kriminalisztikai nyelvész szakértő munkája. In: Belügyi Szemle. vol. 26, p. 93–98 (1988)

Szegedi, Z.: Kriminalisztikai szövegnyelvészet. In: Doktori disszertáció. Budapest (2018)

Szilák, J.: Az írásszokások néhány formai jegyének hátteréről. In: Belügyi Szemle. vol. 16, p. 67–68 (1980)

Tolnainé Kabók, Z.: Interdiszciplináris kapcsolatok a rendészettudományok és az alkalmazott nyelvészet között – különös tekintettel a törvényszéki nyelvészetre. In: Magyar Rendészet. p. 131–145. No. 5 (2015)

Vincze, V.: Uncertainty detection in hungarian texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. p. 1844–1853. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (2014)

Zhang, C., Wu, X., Niu, Z., Ding, W.: Authorship identification from unstructured texts. In: Knowledge-Based Systems. vol. 66, p. 99–111 (2014)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of hungarian. In: Proceedings of RANLP. p. 763–771 (2013)