VINCZE VERONIKA[1] – FŐZŐ ESZTER[2] – KICSI ANDRÁS[3] – VIDÁCS LÁSZLÓ[1,3]

[1]Eötvös Loránd Kutatóhálózat, Mesterséges Intelligencia Kutatócsoport
[2]Nemzetbiztonsági Szakszolgálat Szakértői Intézet
[3]Szegedi Tudományegyetem, Informatikai Intézet
vinczev@inf.u-szeged.hu, fozo.eszter@nbsz.gov.hu, akicsi@inf.u-szeged.hu, lac@inf.u-szeged.hu
https://orcid.org/0000-0002-9844-2194, https://orcid.org/0000-0002-3144-9041,
https://orcid.org/0000-0002-0319-3915

Vincze Veronika – Főző Eszter – Kicsi András – Vidács László: SZIRA: Szövegfeldolgozó Információs Rendszer és Adattár a szerzőazonosítás szolgálatában
Information System and Data Store for Text Processing: Automatic Text Analysis in the Service of Authorship Identification
Alkalmazott Nyelvtudomány, Különszám: Alkalmazott nyelvészet és kriminalisztika, 2022, 52–73.
doi:http://dx.doi.org/10.18460/ANY.K.2022.005

SZIRA: Szövegfeldolgozó Információs Rendszer és Adattár a szerzőazonosítás szolgálatában

Information System and Data Store for Text Processing: Automatic Text Analysis in the Service of Authorship Identification

In this paper, we present SZIRA (Information System and Data Store for Text Processing), which supports the work of linguist experts, aiming at determining whether the author of certain texts is the same or different. The tool is based on automatic linguistic preprocessing of Hungarian texts and is able to extract more than 200 linguistic features belonging to the fields of morphology, syntax, semantics, pragmatics, among others. Based on a small corpus, we also illustrate how SZIRA works in practice.

Keywords: linguistic profile, authorship identification, text processing, automatic analysis, forensic linguistics

1. Bevezetés: a kriminalisztikai szövegnyelvészet

A kriminalisztikai vagy bűnügyi nyelvészet (https://www.iafl.org) az egyéni nyelvhasználat (idiolektus) sajátosságai alapján végez szerzőségvizsgálatot elsősorban a nyomozóhatóságok (elsősorban rendőrség, ügyészség) részére. A szerzőségvizsgálat egyik módszere a **nyelvi profilalkotás/csoportbehatárolás**, mely során a nyelvész szakértő az ismeretlen szerzőjű (kérdéses) írásműből igyekszik feltárni a fogalmazó szocio-demográfiai és szocio-kulturális csoporttulajdonságaira utaló stílusjegyeket, tehát a nemére, életkorára, iskolai végzettségére, anyanyelvére, lakóhelyére stb. vonatozóan von le következtetéseket. A módszer különösen azokban az esetekben alkalmazható, amikor rendelkezésre áll egy vagy több kérdéses szerzőségű (inkriminált) írásmű, az ügyben a nyomozóhatóságoknak még nincs konkrét gyanúsítottja, és a szakértő által adott nyelvi profil alapján próbálják szűkíteni a potenciális elkövetők körét.

A szerzőségvizsgálat másik esete a fogalmazó beazonosítása, a **szerző-azonosítás** (authorship identification, forensic stylistic; Christal et al., 2018); ez

a módszer azokban az esetekben alkalmazható, amikor a nyomozóhatóságnak van(nak) konkrét gyanúsítottja(i). A gyanúsított személy(ek)től beszerzett/felvett szövegminta és az inkriminált írásmű(vek) összehasonlító szövegnyelvészeti elemzésével a szakértő választ adhat arra a kérdésre, hogy a gyanúsított személy fogalmazta-e a kérdéses írásműve(ke)t. Az összehasonlítást megelőzi mind a kérdéses, mind a minta szövegcsoport szeparált elemzése, mely során a szakértő az általános és különös (egyedi) szövegsajátosságokat tárja fel, majd ezeknek a sajátosságoknak az előfordulásait, közvetlen szövegkörnyezetét hasonlítja össze. Az elemzés – így az összehasonlítás is – a nyelv valamennyi rétegén történik (morfológia, szemantika, szintaxis, pragmatika, helyesírás), a szerzőazonosság kimondásához a két szövegcsoport (kérdéses és minta) különös szöveg-sajátosságainak nagyfokú hasonlósága szükséges, és a hasonlóság nem korlátozódhat például csak a szókészletre. Amennyiben mindkét szöveg-csoportban és minden elemzési területen találkozunk következetesen előforduló különös sajátossággal, akkor valószínűsíthető, hogy a két szövegcsoportnak a fogalmazó-szerzője azonos. Ez az egy az egyhez (1 : 1) alapú – vagyis a kérdéses a mintához – hasonlítás a legtöbb kriminalisztikai szakértői gyakorlat része (pl. kézírás, okmány, tárgy, fotó, beszédhang stb.).

Amíg a nyelvi profilalkotásnak meglehetősen szűkösek a szakirodalmi referenciái, addig a szerzőségvizsgálatra irányuló nyelvészet virágzó szakirodalommal rendelkezik elsősorban nemzetközi szintéren és a gépi szöveg-összehasonlítás területén (pl. Rexha et al., 2018; Sousa-Silva 2018; Zhang et al., 2014). Ennek az az oka, hogy a gépi szöveg-összehasonlítás alapja a korpusznyelvészet, mely a mesterséges intelligencia és az NLP-eszközök térnyerésével hatalmas fejlődésnek indult, egyre több a kísérletezés és az eredmény az automatizációt illetően az emberi közreműködés nélküli szövegelemzésre. Nagyszerű gépi elemzők működnek jelenleg is magyar nyelvre (pl. Zsibrita et al., 2013; Simon et al., 2020; Orosz et al., 2022), ám a gépi szövegelemzés és a kriminalisztikai szövegnyelvészet találkozása mindezidáig nem valósult meg. Ennek oka, hogy Magyarországon a nyelvész szakértés intézményesített formában kizárólag az NBSZ Szakértői Intézetben működik, így természetes, hogy a felhasználói igények is innen származnak a megrendelők minél hatékonyabb kiszolgálása érdekében. Az automatizáció szükségességét illetően megerősítenek minket a külföldi partnerszolgálatok tapasztalatai és a nemzetközi kutatási irányok is (Vincze et al., 2021). Az is természetes volt, hogy a gépi szövegfeldolgozással párhuzamosan az igény arra vonatkozóan is megfogalmazódjon – tudomásunk szerint Magyarországon elsőként –, hogy a gépi alkalmazások minimális emberi közreműködéssel képessé váljanak eldönteni két szövegről, hogy a fogalmazójuk azonos-e vagy sem.

Jelen tanulmány célja, hogy bemutasson egy gépi szövegelemző alkalmazást, mely által komplex szöveganalitika valósítható meg: ez a SZIRA, a Nemzetbiztonsági Szakszolgálat (NBSZ) és a Szegedi Tudományegyetem közös kutatásának eredménye, melyet az NBSZ Szakértői Intézet Nyelvész Szakértői Laboratóriumában tervezünk bevezetni.

2. SZIRA

2.1. Bemutatkozás

Az NBSZ Szakértői Intézet Nyelvész Szakértői Laboratóriuma a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Laboratórium Nyelvtechnológiai munkacsoportjában 2020-tól kutatás-fejlesztési projektekben vesz részt. A kutatások a beszédsajátosság életkori változásainak vizsgálatára és a gépi szövegelemzésre irányulnak; utóbbit célzó kutatási projektünk első lépcsőfokaként megvalósult a SZIRA 1.0.0 verziója. Ennek a verziónak a tesztelése, az elemzési képesség bővítése és a pontosság fejlesztése folyamatosan zajlik, hogy a folyamat végén a Szakértői Intézetben bevezetésre kerülhessen a gépi szövegelemzés mint új képesség.

A Szegedi Tudományegyetemmel közösen létrehozott **Szövegfeldolgozó Információs Rendszer és Adattár** (**SZIRA**) korszerű NLP-eszközökre (mesterséges intelligenciára) épülő, kifejezetten magyar nyelvre kifejlesztett gépi szövegelemző rendszer, melyet kriminalisztikai szövegnyelvészeti célból alkalmazunk. A SZIRA – természetesen a digitalizálást követően – bármilyen terjedelmű, kivitelezésű (szövegszerkesztővel, kézírással, írógéppel stb.), műfajú (e-mail, levél, poszt, újságcikk, komment, SMS, IM, hirdetés, hangzó szöveg beszédleirata stb.), különösen fenyegetést, rágalmazást, becsületsértést, zsarolást, csalást megvalósító magyar nyelvű szöveget képes automatikusan feldolgozni és összehasonlítani. A gépi szövegelemzés bevezetésével bizonyos – kriminalisztikai szövegnyelvészeti szempontból jelentős – nyelvi jegyek feltárásához, előfordulásának detektálásához, szövegszerkezeti kalkulációkhoz stb. sokkal gyorsabban jutunk el, ezáltal a gépi szövegelemzők hatékonyan képesek támogatni a nyelvész szakértői munkát, amikor a feladat szövegek/szövegcsoportok nyelvi jellemzőinek statisztikai összevetése az azonos vagy különböző fogalmazó-szerzőség megállapítása céljából.

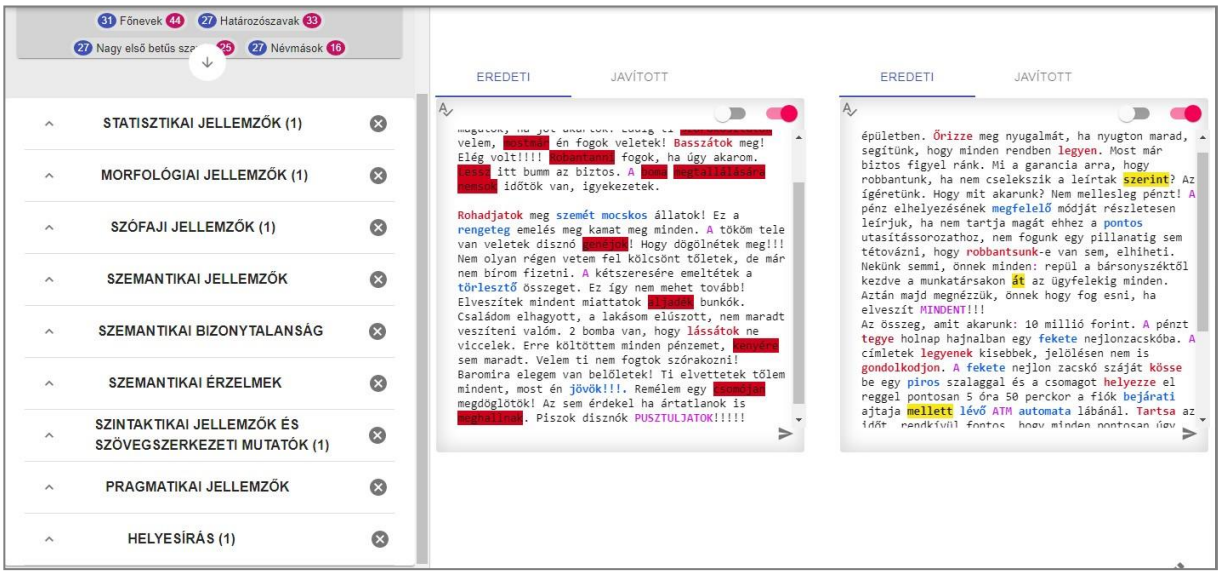
Fontosnak tartjuk hangsúlyozni, hogy a SZIRA létrehozásának célja, hogy a humán szakértő munkavégzését, a véleményalkotás folyamatát gépi szövegelemzéssel és -összehasonlítással, a szerzőazonosság vagy éppen -különbözőség gépi úton történő megbecslésével, valamint NBSZ-specifikus szakmai adattárral támogassa. A SZIRA által automatikusan (és gyorsan) kinyert nyelvi jellemző konkrét szövegbeli megjelenését a szakértő egy kattintással megtekintheti; a gépi szövegelemzéssel bizonyos jellemzőket nem kell többet manuálisan számolni; az automatikus javítás lehetővé teszi az átírás

felgyorsítását; a kinyert diagramok a szakvéleményben illusztrációként szolgálhatnak stb. A SZIRA által végzett elemzések tehát humán nyelvész szakértői kontroll alatt működtethetők, ugyanakkor szélesítik és variábilissá teszik a szakértői elemzési és vizualizációs repertoárt. A SZIRA létrehozásának célja bizonyos manuális szövegelemzési metódus automatizálása, illetve a fogalmazó beazonosítására irányuló kriminalisztikai szövegnyelvészeti vizsgálatok pontosságának növelése.

A SZIRA 1.0.0 verziója 9 kategóriában (statisztika, morfológia, szófaj, szemantika, szemantikai bizonytalanság, szemantikai érzelem, szintaxis, pragmatika, helyesírás) 226 nyelvi jellemző előfordulási száma és aránya, szövegszerkezeti mutató, különféle tulajdonnevek és más entitások (személynév, helységnév, intézménynév, dátum stb.), valamint 8 különböző érzelem (bánat, félelem, düh, meglepődés, nyugtalanság, szeretet, öröm, undor) köré csoportosítható szavak kinyerésére és diagramos összehasonlítására képes. Feltárja a szövegek bizonyos stílusjegyeit, amelyek a fogalmazó-szerző egyedi nyelvhasználati jegyei is egyben; a stilometria tehát segíthet a potenciális elkövető nyelvi profiljának kialakításában (pl. Coulthard, 1994; Michell, 2004). Az érzelmeket és különböző jellegzetes szavakat (pl. agresszió) a SZIRA szólisták alapján azonosítja, bármely szólista szabadon alakítható. Az elemzési eredmény értékei .pdf, .csv és .xlsx kiterjesztésben exportálhatóak. A SZIRA elemző felületén két szöveg/szövegcsoport elemzési értékei jeleníthetők meg egyidejűleg; a felhasználói felület a nyelvész szakértői gyakorlat alapján lett kialakítva, ugyanis leggyakrabban ismeretlen és ismert szerzőségű szövegek összehasonlítása a feladat annak eldöntése érdekében, hogy a két szöveg/szövegcsoport (kérdéses szöveg és mintaszöveg) fogalmazó-szerzője azonos vagy különböző.

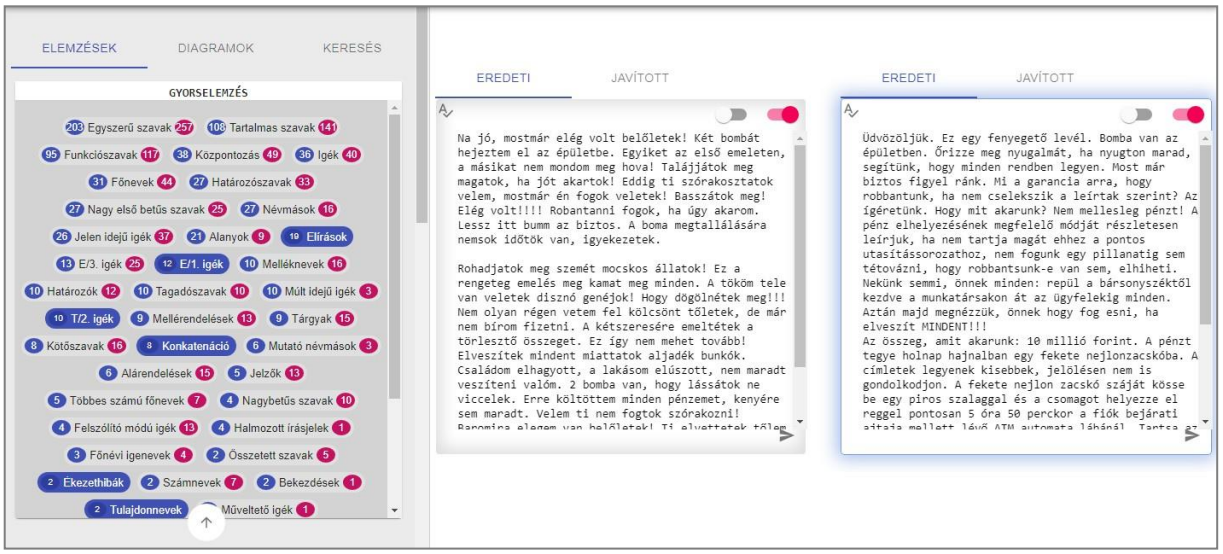
A helytelen írásmódú szóalakok és az egyéb nyelvi jellemzők betű- és háttérszínek különféle kombinációjával, egyéni beállításokkal egyidejűleg jeleníthetők meg a SZIRA elemző felületén. Az 1. ábrán piros háttérszínnel az elírások, sárga háttérszínnel a névutók, piros betűszínnel a felszólító módú igealakok, kék betűszínnel a jelzők, rózsaszín betűszínnel a csupa nagybetűből álló szavak láthatók.

1. ábra. A SZIRA elemző felülete



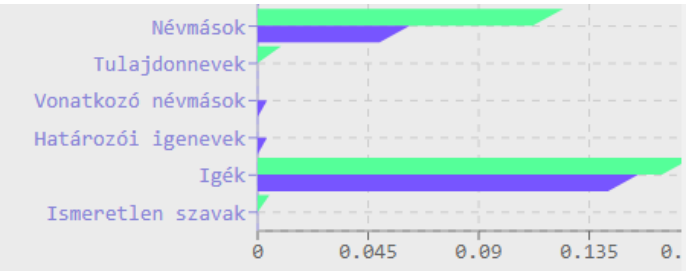
A SZIRA felületén gyorselemzés funkció érhető el, mely gyors összevetési lehetőséget teremt a két szövegre/szövegcsoportra vonatkozóan: a gyorselemzés a szövegbeli előfordulási számadatokat jeleníti meg, a két szöveget/szövegcsoportot különböző színekkel ábrázolja (2. ábra).

2. ábra. Gyorselemzés funkció



Az elemzési eredményekről a SZIRA diagramokat is készít, így az előfordulási számadatok mellett (ld. gyorselemzés) az előfordulási arányok is gyorsan és látványosan összevethetők. A 3. ábrán például az látható, hogy az egyik szövegben közel kétszer annyi névmás fordul elő, mint a másikban.

3. ábra. Diagramos megjelenítés

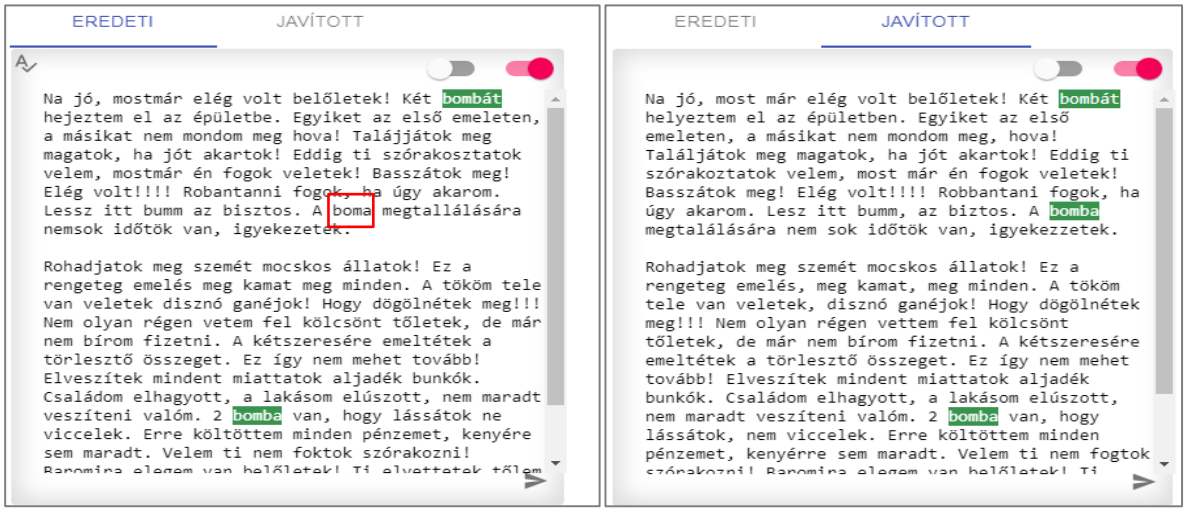


2.2. Helyesírás

A helyesírás kérdése a kriminalisztikai szövegnyelvészetben különösen fontos, nagy bizonyító ereje van a szerzőségvizsgálat során azonosított, következetes típushibáknak. Mivel a bűnügyi nyelvész leggyakrabban nagyon rövid szövegekkel dolgozik, bármely, a szövegből kinyerhető információ jelentős, így például a szövegből kimutatható, a fogalmazóra jellemző kivitelezési eljárások és a szöveg helyesírási és nyelvhelyességi színvonala. Ebből kifolyólag az automatizálást célzó NBSZ-kutatások egyik sarokköve a helyesírás kérdése: a gépi helyesírás-javítás mint képesség bevonása a nyelvész szakértői gyakorlatba.

A helyesírás-ellenőrzés azért olyan fontos a névtelen levelekben, mert a szándékosan vagy nem szándékosan rosszul leírt szóalak a gépi felismerés pontosságát csökkenti, hiszen a rendszer ismeretlen szóként kezeli az adott szóalakot, és elmarad a szó morfológiai, szófaji, szemantikai, szintaktikai, pragmatikai besorolása is. A 4. ábrán látható, hogy a SZIRA kereső funkciójának használatával a *bomba* lemmára kereséskor a rosszul leírt *boma* szóalakot a rendszer nem ismeri fel, azonban a javított szövegváltozatban helyesen azonosítja.

4. ábra. Keresés az eredeti és javított szövegváltozatokban



A SZIRA helyesírás-javító funkciójának célja tehát az, hogy az automatikus helyesírási hiba felismerése és kijavítása megtörténjen. Emellett azonban

kiemelten fontos az is, hogy a rendszerben mind az eredeti, mind a kijavított szövegváltozat egymástól elkülönüljön (ld. a felületen az „Eredeti” és a „Javított” füleket), és a gépi elemző képes legyen külön-külön elemezni mind az eredeti – helytelen, de a fogalmazó egyedi (?) típushibáit tartalmazó –, mind pedig a korrekt szövegváltozatot.

A szöveg helyesírását a SZIRA a Hunspell nevű, széles körben használt helyesírás-ellenőrző, alaktani elemző szoftver segítségével vizsgálja meg, melyet kifejezetten a magyar nyelvhez fejlesztettek ki. Ez a szoftver többféle módszert is alkalmaz annak érdekében, hogy ne csak a helyesírási hibát találja meg, hanem minél jobb ajánlást tegyen a helytelenül írt szó javítására. Olyan bővítést adtunk hozzá a Hunspell-ajánlás keresési folyamatához, amely számításba veszi a többi helyesírási hibát és a lehetséges javítási verziókat, majd közülük kiválasztja a legvalószínűbbet. Ez magába foglalja az ékezethibák és a *j/ly* hibák priorizálását (melyeket egyúttal jellemzőként is használtunk), illetve priorizálja a valószínűleg hibásan egybe- vagy különírt kifejezéseket is.

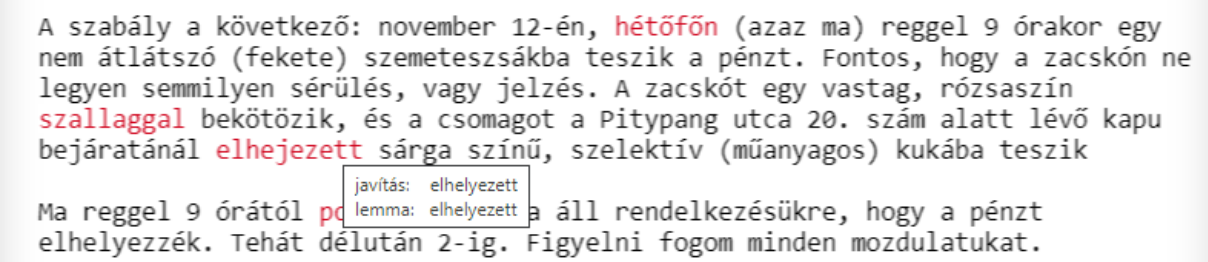
Néhány strukturálatlan, magyar nyelvű, általános témájú, internetes szöveg esetén a szleng kifejezések, rövidítések, idegen nyelvű szavak mellett hangulatjeleket is találhatunk a karakterláncban. Úgy módosítottuk a helyesírási elemzőt, hogy a hangulatjeleket ne vegye figyelembe akkor, amikor helyesírási javítás ajánlása a feladat. Az emotikonok felismeréséhez emotikonkönyvtárat hoztunk létre.

Az elírások javítására több, fejlettebb kísérleti megoldást is kipróbáltunk. Ezek főként a szövegfeldolgozásban a jelenleg az egyik legkorszerűbbnek számító BERT (Devlin et al., 2019) technológián alapuló módszert igyekeztek kiaknázni, ezen belül is a magyar nyelvre fejlesztett HuBERT-et felhasználva (Nemeskey, 2021). Célunk az volt, hogy a kontextustól függő, formailag helyes, de a szövegben mégis helytelenül szereplő szavakat ismerjük fel (pl. *Az ajtót csuklya be Panni, nem az ablakot.*). A módszer nagy mennyiségű tanítóadatot igényel, ezeket magyar Wikipedia-korpuszon állítottuk elő, mivel a Wikipedia az átlagosnál helyesebb (a köznyelvi normához jobban közelítő) szövegezéssel íródik. A tanítóadatokhoz ezekhez mesterséges hibákat szúrtunk be a következők szerint: egy betű hozzáadása egy tetszőleges pozícióra, egy betű kicserélése egy tetszőleges helyen, egy betű kitörlése egy tetszőleges pozícióról, tokenen belül szereplő két különböző betű kicserélése egymással, szóban szereplő „j” kicserélése „ly” betűre, szóban szereplő „ly” kicserélése „j” betűre. Bár a modell pontossága a Hunspell szótáralapú módszerével fel tudja venni a versenyt (kézi kiértékelés alapján 94%), fedése, azaz a detektált hibák mennyisége jelentősen alul múlta a Hunspell teljesítményét. Talált olyan eseteket (főként a rövid szavak körében), ahol valóban kontextustól függően észlelte a Hunspellel ellentétben az elírást, ám ezekre az észlelésekre nem tudtunk általános érvényű szabályt alkotni.

Ezzel kapcsolatban további kísérleteket tervezünk, több kontextusfüggő hiba tesztelésével.

A SZIRA felületén kérhetünk automatikus javítást a rendszertől, de a javítást magunk is elvégezhetjük; ha a rendszer számára nagy mennyiségű az ismeretlen szó (ami lehet rossz helyesírású szó és/vagy a fogalmazó egyedi megoldása), érdemes a szöveg javítását manuálisan elvégezni. A gépi helyesírás-javítást a felületen bármikor felülírhatjuk; ha egy javítással nem értünk egyet, azt szabadon megváltoztathatjuk. A rendszer minden szóra kiírja a hozzá tartozó javítást, a szótövet (lemmát) és megállapított szófajt is (5. ábra). A javítást követően az elemzés újra lefuttatható (hiszen várhatóan eltérő értékek keletkeznek), a rendszer képes rögzíteni mind a kétféle elemzés adta paramétereket egy-egy szöveghez/szövegcsoporthoz.

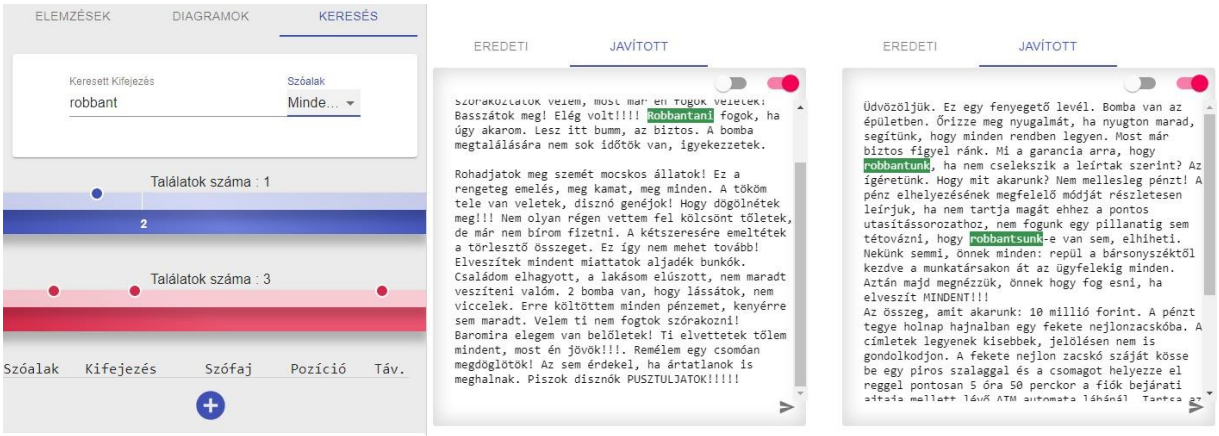
5. ábra. Információk a SZIRA javítási funkciójában



2.3. Részletes kereső

A SZIRA-ban lehetőségünk van konkrét szóalakok, szótövek, szóelemek vagy szókapcsolatok keresésére; a „találatok” pontos helyét a SZIRA a teljes szöveghosszhoz képest is megjeleníti egy diagramon. (6. ábra).

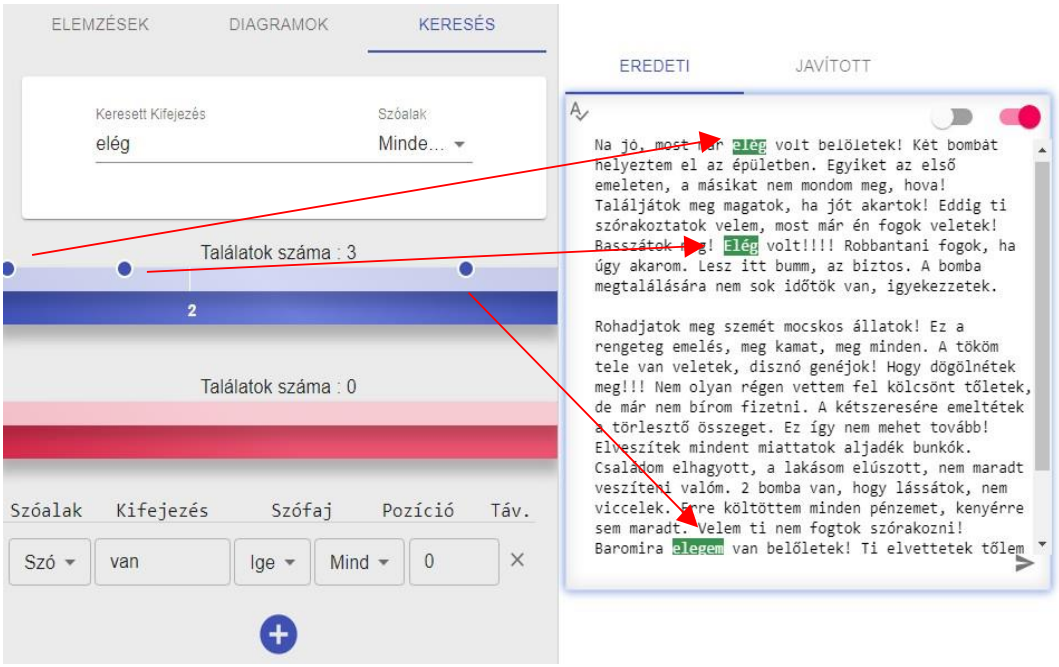
6. ábra. Keresési funkció



A szókapcsolatoknál beállíthatjuk, hogy a két szó milyen távolságra álljon egymástól (az első keresett szóhoz képest a rendszer hány szó távolságon belül keresse a második szót), és a két keresett szó hogyan helyezkedjen el egymáshoz

képest (az első szóhoz képest a második szót előrébb vagy hátrább keresse a rendszer) (7. ábra).

7. ábra. Szókapcsolat-keresés



2.4. Gépi szerzőségvizsgálat

A szerzőegyezés gépi megállapítása tapasztalataink és a tudományos irodalom szerint is nehéz feladat. Kevés valóban korszerű kutatás történt még a témában, a legrelevánsabb hasonló publikációk általában n-gramokkal dolgoznak (Keselj et al., 2003; Stamatatos 2013), de általában ezek is más feladaton, legtöbbször zárt halmazon (változatlan, konstans korpuszon), és jelentősen nagyobb mennyiségű, többnyire szépirodalmi szövegen. Ezek az irodalom szerint 80% körüli pontossággal meg tudják állapítani két szerző egyezését. A szöveg témája is kihat a módszerek pontosságára.

Ha azonban két új szöveget akarnánk összevetni egymással, akkor újra kellene tanítani a modellt, mert ha korábban teljesen ismeretlen szerzőről van szó (ami egy nyomozás során igen valószínű), akkor a korábban előállt tanítás képtelen lenne ezt a szerzőt újabb tanítás nélkül bármely dokumentumhoz hozzárendelni – ez a zárt halmaz nagy hátránya, a tanításkor ismernünk kell az összes lehetséges szerzőt. Ez veszélyes a modell pontosságának fenntartása szempontjából, valamint idő- és erőforrás-igényes feladat; ezért vizsgáltunk más módszereket, más megközelítéseket: az egyszerű n-gram alapút, az elemstabilitás alapút és byte n-gram alapút. Az elemstabilitáson alapuló módszer (Koppel et al., 2006) intuíciója, hogy vannak a nyelvben úgynevezett instabil szavak, kifejezések, amik könnyen helyettesíthetők valami más, ugyanolyan jelentésű kifejezéssel. Ez akár egész tagmondatra vagy teljes mondatra is terjedhet. Ezek meghatározásához

azonban egyrészt nagyobb szövegmennyiségre, másrészt megfelelő alternatív szövegekre lenne szükség. Erre egy jó módszer lehet a gépi fordítás (egy másik nyelvre, például angolra, aztán vissza magyarra), de a tapasztalataink szerint a publikus magyar fordítók még nem elég jó eredményeket adnak ehhez.

A few-shot learning és az active learning témakörében is végeztünk irodalomkutatást, ám végül a feladat szempontjából kevésbé ítéltük őket megfelelőnek. A jelenlegi, nyílt halmazon is általánosan legjobban teljesítő modellünk azonban ennek egy módosított változata. Tanítóhalmaznak a Hunglish 1.0 korpusz magyar részhalmazát használtuk, melyben főként hétköznapi szövegek vannak az index.hu oldaláról. A korpusz elérhető FTP linken keresztül (ftp://mokk.bme.hu/Hunglish). Szűrésünk a legalább 400 karakteres kommenteket hagyta meg, melynek eredményeképp 77018 szövegtörzset használtunk fel. Ez jelentősen nagyobb adathalmaz, mint a rendelkezésünkre álló ügyek vizsgálati anyagait tartalmazó korpusz, és következésképpen sokkal jobban általánosítható. A jelenlegi legjobbnak talált klasszifikáló modell szerzőhasonlóság megállapítására egy BERT architektúrához hasonló, transformer encoder alapú modellt takar. A bemenet először egy szóbeágyazásra alkalmas rétegen megy keresztül. A bemenet tokenizálására, azaz olyan egységekre bontására, melyeket a modell már értelmezni tud, a fentebb említett BERT modell tokenizálóját használjuk. Miután megtörtént a szóbeágyazás, az így készült vektorok több ún. kódoló (encoder) rétegen haladnak keresztül. Az encoder rétegeken való áthaladás után a kapott vektorok egy újabb rétegen haladnak keresztül mielőtt elérik végcéljukat, egy 1 db neuronból álló teljesen kapcsolt réteget. Ezen utolsó réteg felel annak a kérdésnek a megválaszolásáért, hogy két szerző egyezik-e a bemeneti szöveg alapján, avagy sem. A SZIRA jelenlegi pontossága az 1. táblázatban látható (helyes elfogadás = a rendszer helyesen azonosította a fogalmazót; helyes elutasítás = a rendszer helyesen különböztette meg a fogalmazókat; téves elfogadás = a rendszer tévesen azonosította a fogalmazókat; téves elutasítás = a rendszer tévesen különböztette meg a fogalmazókat).

1. táblázat. A SZIRA gépi becslésének pontossága (accuracy)

Helyes elfogadás (True Positive)	Helyes elutasítás (True Negative)	Téves elfogadás (False Positive)	Téves elutasítás (False Negative)
0,40	0,78	0,60	0,22

Ez azt jelenti, hogy a modell jelenleg hajlamos túlbecsülni a szerzők egyezésének valószínűségét. Az egyezés elvetésében viszonylag megbízhatóan működik: a rendszer által különbözőnek vélt fogalmazók 78%-ban valóban nem egyeztek. Természetesen a modell pontosságára nagyban kihat a szövegek hossza is, rövidebb szövegek esetében kevesebb a nyelvi adat, ami alapján a gépi becslés

működik, így nagyobb a hibázás lehetősége is. Célunk a rendszer pontosságának megerősítése, feltételezésünk szerint a pontosság a terjedelemmel arányosan növekszik, nagyobb szövegtörzsek esetén megbízhatóbb eredményt adva. További adatok láthatók az elemzésről a 2. táblázatban. Az eredmények mérésében az egyezőnek tekintett válaszokat vizsgáltuk. A precision érték azt írja le, hogy az egyezőnek kimutatott ügyekből mekkora arány az, amely tényleg egyezést mutat. A fedés érték azt írja le, hogy a valós egyezéseknek mekkora részét fedtük le. Az F-mérték a két érték harmonikus átlaga. Az eredményekből a korábbiaknak megfelelően szintén az látható, hogy a rendszer hajlamos olyan egyezéseket is megállapítani, amelyek nem helyesek. Megjegyezzük, hogy a rendszer egy 0 és 1 közti valós számot produkál minden szövegpárra, amelyet jelen esetben egyszerűen 0.5-nél küszöböltünk. A magas fedés érték szintén inkább azt támasztja alá, hogy a módszer jelentősen jobb lehet az egyezések kizárásában, mint megállapításában.

2. táblázat. A SZIRA gépi becslésének pontossága (precizitás, fedés, F-mérték)

Pontosság (Precision)	**Fedés (Recall)**	**F-mérték (F-score)**
0,2455	0,9718	0,3920

A nemzetközi irodalomban szerzők gépi úton történő összehasonlítása zárt halmazon (authorship attribution, pl. Badirli et al., 2019; Juola, 2008; Sarwar et al., 2018; Qian et al., 2017) viszonylag megbízhatóan működik nagy terjedelmű szövegek összevetése esetén. Az NBSZ nyelvész szakértőinek a feladata azonban a nyílt szöveghalmazon működést követeli meg, hiszen a rendszernek egy új vizsgálati anyagot a már korábban feldolgozott vizsgálati anyagokkal kell összevetnie, vagyis a modellnek nem csupán azokat a szövegeket kell tudnia szerzőkhöz társítani, amelyek szerzőjét már a tanítás során látta, hanem olyan szövegeket is kell tudnia párosítani, amelyek egy teljesen ismeretlen szerzőtől származnak. A másik akadály az, hogy a fenyegető szövegek majdnem mindig igen rövidek, így sokkal kevesebb adatot szolgáltatnak a megalapozott, gépi összevetéshez.

További eredményeink bemutatása során a szerzőazonosság gépi becslésétől jelenleg független jellemzőket mutatunk be. A nyelvész szakértők ezeket ugyan eredményesen fel tudják használni munkájuk során, gépi tanulási kísérleteink azt mutatták, hogy a rendelkezésre álló korpuszon nem találunk megfelelő súlyozást, ami szerint a jellemzők egyértelműen segítenék a modellek munkáját. Önmagukban, a szöveg felhasználása nélkül nem találtunk olyan súlymátrixot, ami egyértelműen jó indikátora lett volna a szerzők egyezésének. Fontos megjegyezni, hogy itt nem csupán az alább, a kísérletek során bemutatásra kerülő szerzők szövegein végeztük a vizsgálatot, hanem a rendelkezésünkre álló összes szövegen. Kis számú szerző mellett természetesen jól kimutathatók a jellemzők

utalása egy bizonyos szerzőre, de ez már a zárt halmazos szerzőazonosítás kérdése lenne, amely a korábban említettek alapján jelentősen könnyebb megoldást kínálna. Érdemes további kutatásokat végezni ebben az irányban, a probléma jelenleg világszinten megoldatlan, ezért a magasabb pontosság elérése komoly feladat.

2.5. Adattár

Az adatbázis célja az elkövető-egybeesés felderítése. Ehhez a szakmai adattár az ügy alapadatainak tárolására, az adatokban történő célirányos keresésre, valamint a vizsgálati anyagokon elvégzett gépi elemzési értékek eltárolására, összehasonlítására alkalmas nyilvántartó rendszer. Az adattárban adott ügyhöz kapcsolódóan a kérdéses írásművek és a szövegminták fogalmazójának profiladatai is tárolódnak (személyes adatok nélkül), valamint maguk a vizsgálati anyagok eredeti és javított szövegváltozatai.

3. Módszer, vizsgálati anyag

Az interneten és a mobileszközökön elképesztő mennyiségű szöveges dokumentum keletkezik; a kommunikáció írott formája mára szinte teljes mértékben az online térre korlátozódik. A 21. századi kommunikációra jellemző, hogy független helytől és/vagy időtől, az emberek állandó elérhetőségére és a közöttük lévő folyamatos információ-megosztásra épül. Az emberek szívesen alkotnak véleményt az online térben kommentekben, posztokban, fejezik ki nézeteiket különböző fórumokon. A modern technológiák tehát elősegítik a személyek közötti kapcsolattartást, és megkönnyítik az emberek véleményének közlését, üzenetek küldését, ugyanakkor az online platformok lehetőséget is teremtenek a szerző valódi kilétének elrejtésére, lehetővé téve anonim üzenetek küldését mások zsarolására, rágalmazására, fenyegetésére vagy csalás elkövetésére.

Az online környezetben megvalósított bűncselekmények elkövetői a legtöbbször lenyomozhatók az IT-szakterület által a digitális ujjlenyomatuk alapján; ilyenkor a nyelvész szakértő bevonása az inkriminált szövegek fogalmazójának megállapítása céljából nem feltétlenül szükséges, hiszen munkája időigényesebb, ezért drágább is. Azoknál a bűnelkövetőknél azonban, akik jól rejtik magukat a digitális térben, a nyelvész szakértők jelenthetik a megoldást, ugyanis képesek az egyre nagyobb számú internetes (szóbeli és írott formában elkövetett) bűncselekményekhez köthető, névtelen vagy álnéven megvalósított interakciók szövegnyelvészeti feldolgozására. Mivel a fenyegetés, zaklatás, zsarolás, rágalmazás, becsületsértés, csalás, köz- és magánokirat-hamisítás stb. sok országban bűncselekménynek minősül – Magyarországon ezt a Büntető Törvénykönyvről szóló 2012. évi C. törvény szabályozza –, a bűnüldöző hatóságok munkáját elősegítendő indokolt az ilyen üzenetek, bejegyzések

küldőinek azonosítása.

A modern igazságügyi nyelvészet vizsgálati anyagai e-mailek, posztok, kommentek, IM-üzenetek, cset- és fórumbejegyzések, valamint SMS-ek. Ezek a műfajok rövid terjedelmükkel, információsűrítő tulajdonságukkal (Prószéky, 2017) indokolttá teszik a hagyományos szövegelemzési eljárások felülvizsgálatát, ugyanakkor szerencsére nagy mennyiségben állnak rendelkezésre, így ugyanattól a szerzőtől a szövegfajták széles palettája vonható vizsgálat alá az egyéni stílussajátosságok minél szélesebb körű feltárása érdekében.

A kommunikációs platformok megváltozásával az inkriminált szövegek is megváltoztak, így a hagyományos kriminalisztikai szövegnyelvészeti elemzés sok esetben nem alkalmazható, ami kihívás elé állítja szakértőt. Az olyan hagyományos elemzési szempontok, mint a helyesírás vizsgálata, jelentősen korlátozottá válik, hiszen például, egy ékezet nélkül írott szövegben az esetleges ékezethibák nem térképezhetők fel. A szövegszerkesztő alkalmazásokban az automatikus helyesírás-javító funkció felülírja a fogalmazó hibáit; az IM-üzenetekben a különírás-egybeírás, a kis és nagy kezdőbetűk, az írásjelek stb. irrelevánsak; a mobileszközökön működő automatikus szófelajánlás nem a fogalmazó szókincsét tükrözi és így tovább. Vagyis az idiolektus részének tekinthető nyelvi jellemzők egy része nem nyerhető ki a szövegből. A megoldás: másfajta szöveg-feldolgozási szempontokkal kell kísérletezni.

Jelen vizsgálat elvégzéséhez hét olyan ügyet választottunk ki, melyben nyelvész szakvélemény készült. Mindegyik ügyben a szakértői vizsgálatok szerzőazonosításra irányultak – tehát voltak az ügyekben kérdéses és összehasonlító írásművek is – és az ügyet eredményesen lezárta a Rendőrség; ezáltal egyrészt ismertté váltak számunkra a fogalmazói csoporttulajdonságok, másrészt ugyanattól a személytől rendelkezésre állt inkriminált írásmű és szövegminta is, így a névtelen szerzőség stílusjegyeit is vizsgálhattuk, például azt, hogy hogyan fogalmaz adott személy, ha névtelenül ír: megpróbálja-e stílusát torzítani, és ha igen, hogyan. A választott hét ügyben hét különböző elkövetőről beszélhetünk, így az esetleges stílusbeli hasonlóságok nem a fogalmazók azonosságára, hanem a fogalmazói csoportjegyek hasonlóságára utalnak. A mostani vizsgálat a szándékos/nem szándékos torzítás nyelvi jegyeinek részletezésére nem tér ki; jelen vizsgálat célja, hogy a hét szövegcsoport gépi szövegelemzési értékei által a fogalmazói idiolektusok különbözőségét szemléltessük.

A 2.5. pontban részletezett online szövegek feldolgozásával szembeni kihívás miatt jelen vizsgálat során törekedtünk olyan szövegekkel dolgozni, melyek online környezetben keletkeztek. Olyan bűncselekmény vizsgálati anyagjain folytattunk elemzéseket, melyekben az inkriminált írásművek és a szövegminták műfajukat illetően elsősorban posztok, e-mailek, rövid üzenetek (azonnali üzenet: IM – Instant Message –, pl. Facebook Messenger vagy SMS), másodsorban

szövegszerkesztővel készült levelek, szórólapok, esetleg hivatalos iratok voltak.

A hét ügyből két esetben közveszéllyel – bombarobbantással – fenyegetés, két esetben nagy nyilvánosság előtt elkövetett becsületsértés és/vagy rágalmazás, két esetben zaklatás, valamint egy ízben hamis tanúzásra felhívás vétsége miatt indult eljárás. A bűncselekmények kategóriáiból is látszik, hogy az ügyekben a téma különböző, ugyanakkor a szókészlet részletes vizsgálata megvalósítható, ti. a tartalmas szavak tekintetében nem beszélhetünk nagyfokú egybeesésről.

A választott hét ügyben a fogalmazói csoporttulajdonságok hasonlóak egymáshoz, nem véletlenül: a kiválasztásnál fontos szempont volt, hogy a fogalmazók közel azonos csoporttulajdonságokkal rendelkezzenek, így a nyelvhasználatukban jelentkező esetleges különbségek egyedi jellemzőknek (idiolektikus elemeknek) minősülnek. A hét ügyben valamennyi fogalmazó férfi, mindannyian középkorúak (életkoruk 30 és 60 év közötti), öten középszintű iskolai végzettséggel (érettségivel), ketten pedig főiskolai/egyetemi diplomával rendelkeznek. A fogalmazók csoporttulajdonságait a 3. táblázat szemléleti.

3. táblázat. A fogalmazók csoporttulajdonságai

Fogalmazó	Nem	Életkor	Iskolai végzettség
A	férfi	38 éves	érettségi
B	férfi	47 éves	érettségi
C	férfi	52 éves	érettségi
D	férfi	42 éves	érettségi
E	férfi	30 éves	érettségi
F	férfi	30 éves	diploma
G	férfi	53 éves	diploma

A korpusz nem kiegyensúlyozott az egyes fogalmazókhoz tartozó szövegmennyiség tekintetében (4. táblázat). Ha az átlagot nézzük, egy-egy fogalmazóhoz 9052 inkriminált token és 16278 szövegminta-token tartozik (a számadatok az eredeti szövegre vonatkoznak, nem a SZIRA által automatikusan javított változatra).

4. táblázat. A korpusz mérete

Fogalmazó	Tokenszám		
	Kérdéses írásművek	**Szövegminták**	**Összesen**
A	440	1577	2017
B	4322	44329	48651
C	489	6271	6760
D	974	7116	8090
E	15016	7883	22899
F	6823	27410	34233
G	35210	19359	54569
	63364	**113945**	**177309**

A vizsgálati anyagokat a SZIRA-val elemeztük: integrált magyarlanc nyelvi elemzővel (Zsibrita et al., 2013) 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,
- összetett szavak száma és aránya.

Szintaktikai jellemzők:

- alanyok, tárgyak, jelzők, határozók, alárendelő és mellérendelő mondatok 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,
- agresszív, 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-Vincze, 2016) utaló szavak száma és aránya.

Pragmatikai jellemzők:

- beszédaktusok száma és aránya,
- idézőjelek és gondolatjelek száma és aránya,
- kifelé és befelé forduló igék száma és aránya,
- meggyőzést jelentő igék száma és aránya,
- diskurzusjelölők száma és aránya.

Helyesírási hibák:

- elírások száma és aránya,
- ékezethibák száma és aránya,
- j-ly hibák száma és aránya,
- konkatenáció (egybeírás-különírási hibák) száma és aránya.

Habár a korszerű magyar nyelvészetben van néhány újabb nyelvi elemző, amely egyes funkciókban jobb eredményeket ígér a magyarlanc-nál, ezek felmérésünk alapján nem támogatnak minden funkciót, amelyre az elemzés során szükségünk volt. A modern trendek alapján például a konstituens-elemzés, melyre rendszerünk több esetben épít, általában nem támogatott.

Jelen vizsgálatban A–G fogalmazók stílusának összevetését kizárólag a kinyert nyelvi jellemzők alapján végeztük el, a gépi becslés eredményét figyelmen kívül hagytuk.

4. Eredmények

Az alábbiakban csak a legfontosabbakra koncentrálva mutatjuk be vizsgálataink eredményét. Mivel a szerzőktől eltérő nagyságú szövegmennyiség állt rendelkezésünkre, elsősorban a nyelvi jelenségek arányaira összpontosítunk, nem a darabszám szerinti előfordulásra, összevetve a gyakorisági adatokat a kérdéses írásművekben és a szövegmintákban is. Az eredményeket az 5–6. táblázatok szemléltetik, továbbá a 8–11. ábrák az egyes fogalmazók szövegeiben mutatják a gyakoriságokat (bal oldalon kérdéses szövegek, jobb oldalon a mintaszövegek alapján).

5. táblázat. A morfoszintaktikai jellemzők átlagos gyakorisága

	A határozószavak aránya		**Az összetett szavak aránya**		**A kötőszavak aránya**		**Az alárendelések aránya**	
Fogalmazó	Kérdéses	Minta	Kérdéses	Minta	Kérdéses	Minta	Kérdéses	Minta
A	0,0727	0,0765	0,0432	0,0398	0,0795	0,0568	0,0386	0,0337
B	0,0572	0,0856	0,0240	0,0182	0,0942	0,0800	0,0365	0,0383
C	0,0511	0,0462	0,0164	0,0235	0,1022	0,0573	0,0593	0,0354
D	0,1488	0,1583	0,0222	0,0055	0,1093	0,0860	0,0657	0,0453
E	0,1305	0,1522	0,0152	0,0104	0,0992	0,1048	0,0629	0,0640
F	0,0637	0,0419	0,0365	0,0380	0,0592	0,0540	0,0273	0,0291
G	0,1022	0,0836	0,0212	0,0394	0,0706	0,0678	0,0411	0,0345

6. táblázat. Statisztikai, szemantikai és pragmatikai jellemzők átlagos gyakorisága

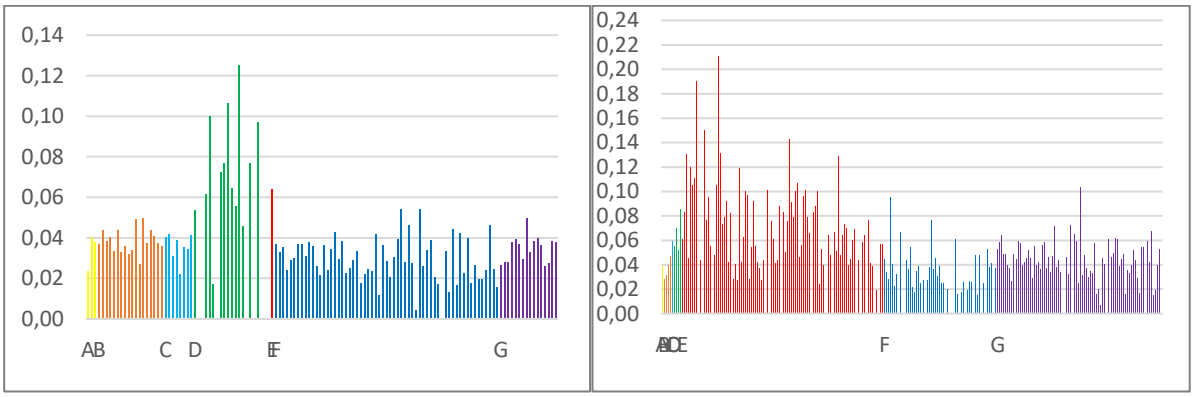
	Felszólító mondatok aránya		Többszörös mondatvégi írásjelek aránya		Megszólítások aránya		Diskurzusjelölők aránya	
Fogalmazó	Kérdéses	Minta	Kérdéses	Minta	Kérdéses	Minta	Kérdéses	Minta
A	0,0000	0,0909	0,0000	0,0423	0,0000	0,0000	0,0250	0,0341
B	0,8182	0,6682	0,0146	0,0155	0,0000	0,0000	0,0334	0,0434
C	0,0500	0,0348	0,0000	0,0013	0,0000	0,0011	0,0532	0,0241
D	0,5131	0,6569	0,1472	0,4423	0,0006	0,0000	0,0535	0,0836
E	0,6605	0,4129	0,2156	0,1665	0,0001	0,0000	0,0921	0,1035
F	0,2201	0,1111	0,0405	0,0023	0,0003	0,0005	0,0327	0,0246
G	0,2886	0,2040	0,0418	0,0048	0,0007	0,0005	0,0448	0,0421

Megállapítható, hogy apróbb eltérések azonos fogalmazó esetében is vannak a kérdéses és mintaszövegek gyakorisági értékeiben (piros betűszínnel jelölve); éppen ezért kell a szöveget komplex elemzés alá vonni, és valamennyi nyelvi szinten lefolytatni az összehasonlító vizsgálatokat, hiszen egy-egy nyelvi jellemzőt érintő eltérés – az ugyanazon szerzőhöz tartozó inkriminált és mintaszöveg között – hibás következtetés levonását vonhatná maga után. Mindazonáltal az 5–6. táblázatból a fogalmazók közötti különbségek is kirajzolódnak. Ha a táblázatok kérdéses és minta oszlopait összevetjük egymással

minden vizsgált jellemzőre nézve, azt láthatjuk, hogy általánosságban hasonló arányban szerepelnek mind a kérdéses, mind a mintaként rendelkezésre álló szövegekben az adott nyelvi jegyek.

Az adott nyelvi jegy előfordulási következetessége a 8–11. ábrákon látszik jobban, ahol a grafikon vízszintes sorában valamennyi fogalmazó valamennyi vizsgálati anyagát megjelenítettük (a fogalmazók eltérő színnel lettek jelölve).

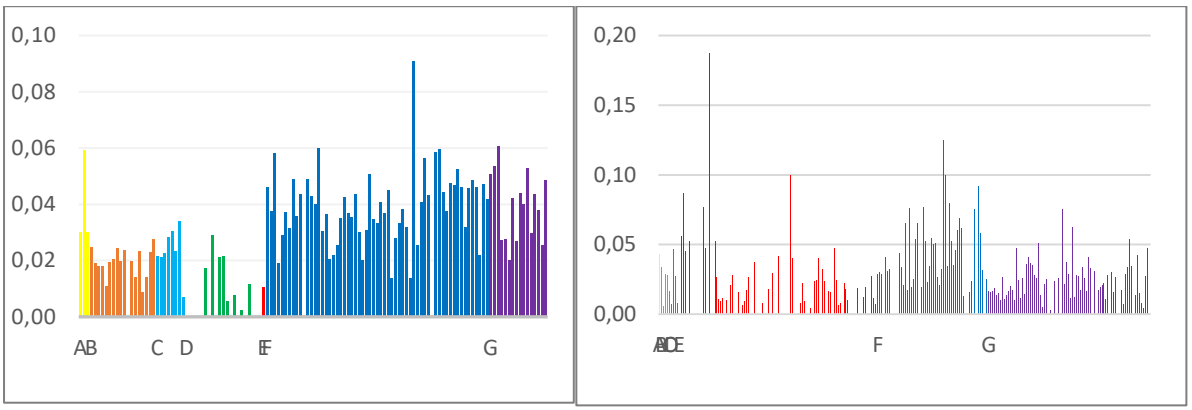
8. ábra. Az alárendelések gyakorisága



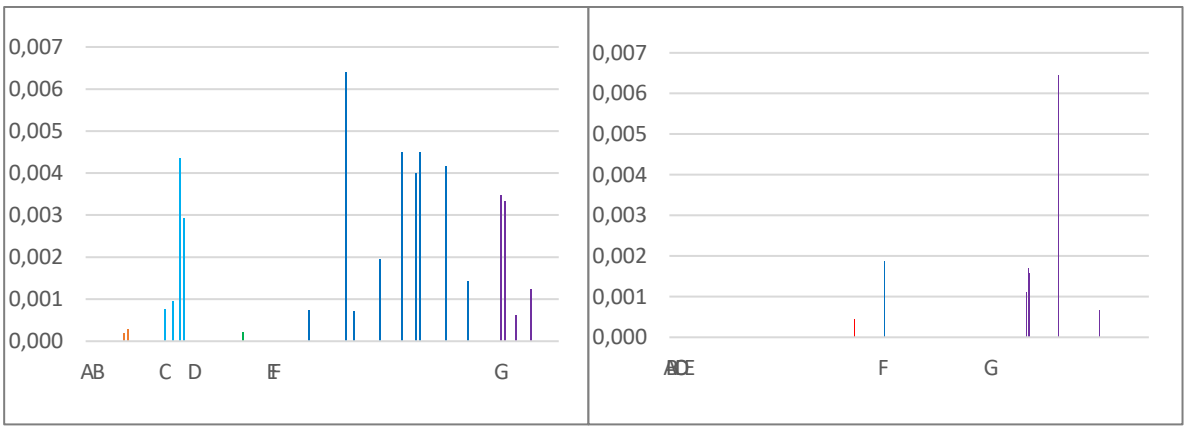
Konkrétabban vizsgálva az egyes jellemzőket: a morfoszintaktikai jellemzőkkel kapcsolatban elmondható, hogy a D és E személy jóval gyakrabban használ határozószavakat (a teljes szószámhoz viszonyítva), mint a többi személy, míg a C és F személy az átlagosnál kevesebbszer használja ezeket. Az E személy az átlagnál gyakrabban használ kötőszavakat és alárendeléseket is (8. ábra), valószínűleg az ő szövegeikre a valamivel bonyolultabb mondatszerkesztés jellemző.

Az összetett szavak vizsgálata is érdekes különbségekre mutat rá a vizsgált személyek között (9. ábra): a D személy kevesebb összetett szót használ, pontosabban a morfológiai elemző azonosít kevesebb összetett szót az általa írt szövegekben. Ez a tény helyesírási hibákra is rámutathat: D személy nincs teljesen tisztában az összetett szavak helyesírására vonatkozó szabályokkal, így – helytelenül – külön is írhatja az egybeírandó szavakat. Ez pedig végső soron az iskolázottsági szintjével állhat összefüggésben. Mivel ezt a feltevést a helyesírási elemzés nem támasztotta alá, nagyobb a valószínűsége annak, hogy az összetett szavak kerülése mögött a műfaji sajátosságok állnak: D fogalmazó esetében a szövegminták jelentős része IM keretében jött létre. A beálló melléknévi igenevek használata is csak néhány szerző (F–G) stílusára jellemző, ami szintén árulkodó lehet a szerzőazonosításra nézve (10. ábra).

9. ábra. Az összetett szavak gyakorisága

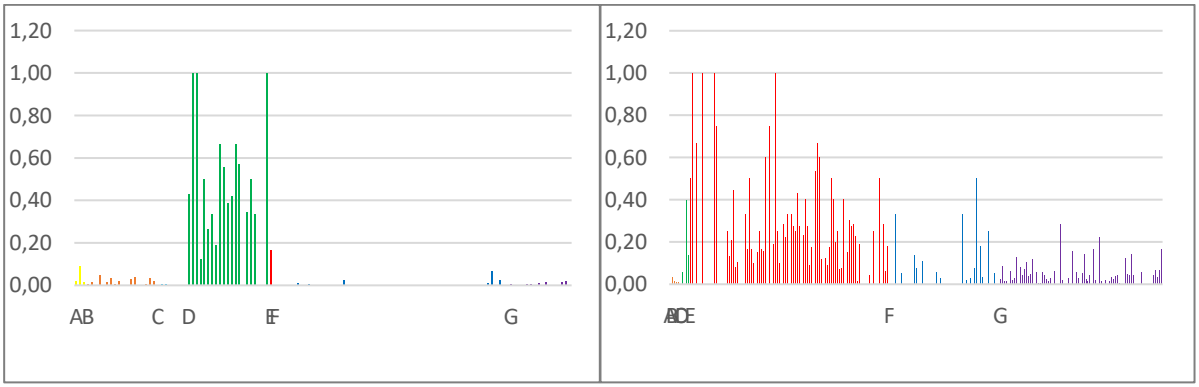


10. ábra. A beálló melléknévi igenevek gyakorisága



Az írásjelek és mondattípusok vizsgálata is érdekességekre mutat rá. Látszik, hogy néhány személy rendkívül magas számban alkalmaz felszólító mondatokat a fogalmazványaiban (B, D, E), míg más fogalmazóknál alig fordul elő ez a mondattípus. Így akár a felszólító mondatok aránya is árulkodó lehet a fogalmazó személyét illetően. Hasonló viselkedést mutat a többszörös írásjelek *(?!, ??, !!! stb.)* használata is: a D és E személyek előszeretettel alkalmazzák ezeket, míg más fogalmazóknál nem figyelhető meg ez a tendencia (11. ábra).

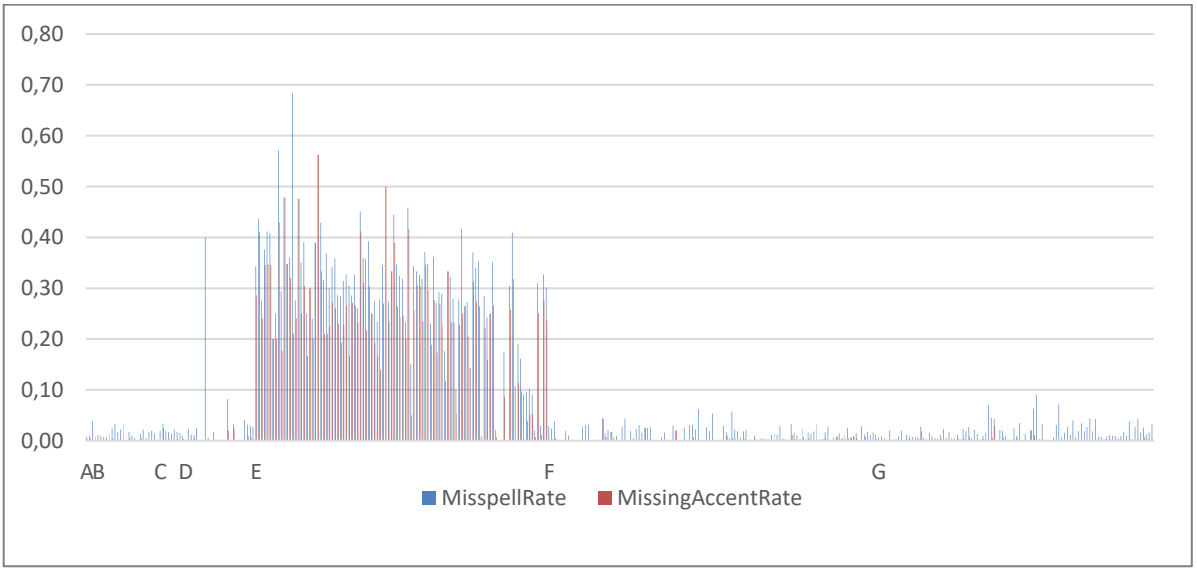
11. ábra. A többszörös írásjelek gyakorisága



Néhány szemantikai-pragmatikai jellemzőt vizsgálat alá véve megállapíthatjuk, hogy szintén az egyéni nyelvhasználatot tükrözik. A megszólítások és záró formulák egyes fogalmazóknál egyáltalán nem, vagy csak minimálisan jelennek meg, míg a F és G személyek látszólag következetesen használják ezeket, nyelvi stílusuk egyik fontos jellemzőjeként. A diskurzusjelölők használata is jellemző az egyénre: a D és E személyek az átlagnál gyakrabban használják őket fogalmazványaikban, mint a többi vizsgált személy.

A helyesírás kategóriáján belül a két legjellemzőbb hibaforrást tekintve – elírás és ékezethiba – azt tapasztaltuk, hogy a korpusz fogalmazóinál az elírások aránya 5% alatt mozog, az ékezethibák aránya még ennél is alacsonyabb (12. ábra). Az E szerzőnél ugyanakkor kiugróan magas mindkét érték. A hiányzó ékezetek magas számát nem feltétlenül az ékezethasználati készség elégtelensége okozza: az írásművekből feltehetően az angol billentyűzet miatt hiányoznak a magyar ékezetek. Mindamellett az E fogalmazóra jellemző nagy számú elírásra (betűcsere, betűhiány, felesleges betű stb.) nincs ésszerű magyarázat, azt gyanítjuk, a szövegtípusnak lehet hozzá köze, ti. mind a kérdéses, mind a mintaszövegek azonnali üzenetküldő szolgáltatáson keletkezett szövegek – Instant Messaging – voltak, mely szövegtípus a prediktív szövegbeviteli opció miatt nagyobb arányban tartalmaz ilyen jellegű hibát (pl. Érsok, 2004). A 12. ábrán feltüntettük a kérdéses és mintaszövegek elírásainak és ékezethibáinak arányát is, hogy láthatóvá váljon, ezen helyesírási hibák előfordulási aránya közel azonos adott fogalmazónál akkor is, ha névtelenül ír (kérdéses), és akkor is, ha névvel (minta), vagyis a helyesírási hibák jellege és előfordulási aránya – természetesen a szövegtípus hasonlósága mellett – általában jellemzi a fogalmazókat, a szerzőségvizsgálat műveletében egyedi jegynek értékelhető.

12. ábra. Az elírások és ékezethibák gyakorisága



5. Összegzés

E tanulmányban bemutattuk a Nemzetbiztonsági Szakszolgálat Szakértői Intézet Nyelvész Szakértői Laboratóriumában kísérleti jelleggel alkalmazott Szövegfeldolgozó Információs Rendszer és Adattárat, a SZIRA-t. A SZIRA működését példákkal illusztráltuk, valamint rámutattunk, hogy a kinyerhető adatok segítségével a SZIRA létrehozásának legfőbb célja megvalósult: kriminalisztikai szempontú gépi szövegelemző alkalmazásként segíti a szövegek/szövegcsoportok összehasonlító nyelvész szakértői vizsgálatát.

Köszönetnyilvánítás

A publikációban szereplő kutatás az Európai Unió támogatásával valósult meg, az RRF-2.3.1-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében.

Irodalom

Badirli, S., Borgo Ton, M., Gungor, A. & Dundar, M. (2019). *Open Set Authorship Attribution toward Demystifying Victorian Periodicals*. Letöltés: https://arxiv.org/abs/1912.08259

Christal, G., Manve, P., Ahuja, P. & Dahiya, M. S. (2018). Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. *Journal of Forensic Science and Criminal Investigation, 8,* 1–10. doi:10.19080/JFSCI.2018.08.555733

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics, 25*. 431–447. doi:10.1093/applin/25.4.431

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference* 4171–4186.

Érsok Nikoletta Ágnes (2004). Sömös, susmus, írj vissza. *Magyar Nyelvőr, 128,* 294–313. Letöltés: http://c3.hu/nyelvor/period/1283/128303.pdf

Hunglish korpusz (2022. 04. 10.), Letöltés: http://mokk.bme.hu/resources/hunglishcorpus/

Juola, P. (2008). Authorship Attribution. *Foundations and Trends in Information Retrieval, 1(3),* 233–334. doi:10.1561/1500000005

Keselj, V., Peng, F., Cercone, N. & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics* 255–264.

Koppel, M., Akiva, N. & Dagan, I. (2006): Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology, 57,* 1519–1525.

Michell, C. S. (2013). *Investigating the use of forensic stylistic and stylometric techniques in the analyses of authorship on a publicly accessible social networking site (Facebook).* Letöltés: https://lexically.net/wordsmith/corpus_linguistics_links/Michell_Facebook_dissertation.pdf

Nemeskey Dávid Márk (2021). Introducing huBERT, In Berend Gábor, Gosztolya Gábor & Vincze Veronika (szerk.), *XVII. Magyar Számítógépes Nyelvészeti Konferencia* (3–14). Szeged: SZTE. Letöltés: https://rgai.inf.u-szeged.hu/sites/rgai.inf.u-szeged.hu/files/mszny2021.pdf

Orosz György, Szántó Zsolt, Berkecz Péter, Szabó Gergő, Farkas Richárd (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In Berend Gábor, Gosztolya Gábor & Vincze Veronika (szerk.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia* (59–73). Szeged: SZTE.

Prószéky Gábor (2017). A számítógép, az elektronikus kommunikáció és az internet hatása. In Tolcsvai Nagy Gábor (szerk.), *A magyar nyelv jelene és jövője* (321–335) Budapest: Gondolat Kiadó.

Qian, C., He, T. & Zhang, R. (2017). *Deep Learning based Authorship Identification*. Letöltés: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf

Rexha, A., Kröll, M., Ziak, H. & Kern, R. (2018). Authorship identification of documents with high content similarity. *Scientometrics*, *115(1)*, 223–237. doi:10.1007/s11192-018-2661-6

Sarwar, R., Yu, C., Tungare, N., Chitavisutthivong, K., Sriratanawilai, S., Xu, Y., Chow, D., Rakthanmanon, T. & Nutanong, S. (2018). An Effective and Scalable Framework for Authorship Attribution Query Processing. *IEEE Access*, *6*, 50030–50048. doi:10.1109/ACCESS.2018.2869198

Simon Eszter, Indig Balázs, Kalivoda Ágnes, Mittelholcz Iván, Sass Bálint & Vadász Noémi (2020). Újabb fejlemények az e-magyar háza táján. In Berend Gábor, Gosztolya Gábor & Vincze Veronika (szerk.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia* (29–42). Szeged: SZTE.

Sousa-Silva, R. (2018). *Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts.* Letöltés: https://www.researchgate.net/publication/333582921_Computational_Forensic_Linguistics_An_Overview_of_Computational_Applications_in_Forensic_Contexts

Stamatatos, E. (2013). On the Robustness of Authorship Attribution Based on Character N-Gram Features. *Journal of Law & Policy*. Vol. 21 (2). Letöltés: https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/7/

Szabó Martina Katalin (2015). Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In Gecső Tamás (szerk.), *Nyelv, kultúra, társadalom* (278– 285). Székesfehérvár, Budapest: Kodolányi János Főiskola, Tinta Könyvkiadó.

Szabó Martina Katalin & Vincze Veronika (2016). Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In Reményi Andrea Ágnes, Sárdi Csilla & Tóth Zsuzsa (szerk.), *Távlatok a mai magyar alkalmazott nyelvészetben* (282–292). Budapest: Tinta Könyvkiadó.

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

Vincze Veronika, Kicsi András, Főző Eszter & Vidács László (2021). A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei. In Berend Gábor, Gosztolya Gábor & Vincze Veronika (szerk.), *XVII. Magyar Számítógépes Nyelvészeti Konferencia* (275–288). Szeged: SZTE.

Zhang, C., Wu, X., Niu, Z. & Ding, W. (2014). Authorship identification from unstructured texts. *Knowledge-Based Systems*, *66*, 99–111. doi:10.1016/j.knosys.2014.04.025

Zsibrita János, Vincze Veronika & Farkas Richárd (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* 763–771. Shoumen: INCOMA Ltd.