Computer-Aided Forensic Authorship Identification in Criminology

András Kicsi[1,2][0000−0002−3144−9041], Péter Sánta[2], Dániel Horváth[2], Norbert Kőhegyi[2], Viktor Szvoreny[2], Veronika Vincze[1][0000−0002−9844−2194], Eszter Főző[3], and László Vidács[1,2][0000−0002−0319−3915]

[1]MTA-SZTE Research Group on Artificial Intelligence
6720 Szeged, Tisza Lajos krt. 103., Hungary 103.
[2]University of Szeged, Department of Software Engineering
6720 Szeged Dugonics tér 13. Hungary
[3]Special Service for National Security
Budapest, Törökvész út 32-34., Bég u., 1022
{akicsi,speter,szvoreny,hoda,knorbert,vinczev,lac}@inf.u-szeged.hu
fozo.eszter@nbsz.gov.hu

Abstract. The increasingly anonymous methods people use to communicate in the modern world allow for more freedom of speech. The safety of anonymity, however, can enable criminals to cause harm to others through various means, such as blackmail, verbal abuse, threat letters and numerous other ways. These culprits, often hiding behind computer screens, can be extremely difficult to identify and especially difficult to find definitive proof of their wrongdoings. They are not completely untraceable, however, as they are bound to leave clues in the text, linking it to them. The way they phrase sentences, the words they use, how often they use them and other parts of their idiolect can be used to identify them and even connect them to other texts. Through analyzing the text, it becomes possible to catch these individuals. This analysis is neither simple nor cheap, the aid of linguistic experts is critical, and even they are likely to encounter difficulties. This article explores the way in which the work of such experts can be assisted through computer analysis based on machine learning techniques and the role Artificial Intelligence plays in bringing these criminals to justice. Our current paper investigates how linguistic features can be automatically extracted to be used in the field. Through a total of 61 real text artefacts written in the Hungarian language by four different individuals, we extract various syntactic and semantic linguistic features which reflect the author's idiolect and aid the expert's work. We demonstrate how the technique can aid author identification in criminology.

Keywords: author identification · stylometry · criminalistics · nlp

1 Introduction

When people talk on the Internet, they generally don't worry too much about the consequences of their comments, they feel safe since they are anonymous. On the one hand, this allows people to talk freely, be honest, to not worry about what others might think of them. On the other hand, this same safety can be exploited in dangerous ways. In the modern world, people are sending text messages anonymously all the time, be it through a forum, a chat room, or even through e-mail with a fake address. This is generally harmless, however, sometimes, things can go very wrong: what people can do with the anonymity they have ranges from lighter wrongdoings such as verbal abuse and deception to very severe crimes such as blackmail, stalking messages or even death threats. The thought of some person one doesn't even know stalking, or perhaps even threatening to harm an individual, can be quite frightening. Perhaps the better question to ask here, however, is not the "is it possible", but rather "how it is possible" to identify the authors of such a malicious message. In criminology, there are linguistic experts who are tasked with solving this problem. Their job is rather difficult and expensive, but it is very important, aiding them could be a worthwhile effort.

There are many features one can use to describe a text. For example, one can analyze how many words there are in a text, how many sentences, or perhaps the average length of these sentences. One can also investigate a text's semantics. For example, the text might be using overwhelmingly negative or positive words, or maybe it contains a lot of racist or aggressive remarks. These features, among others, are all parts of an author's idiolect, and many of them can be used to potentially get a hint about the writer's identity. While some of these are obvious indicators of an author's identity -such as the number of aggressive words they use-, some give more subtle clues about the writer: for example, somebody who uses complex sentences could be said to prefer writing in a live speech style [20].

These features -although very useful- can prove to take a lot of time and effort to extract from the text, and for this reason, automation of this process is crucial. This is where modern technology, equipped with artificial intelligence, can play a significant part. Using technology based on machine learning, we can not only speed up the process of said feature extraction, but we can also take a lot of burden off of the experts so that they can focus more on actually utilizing said features. Of course, it is critical for said systems to be as accurate as they can possibly be, in order for them to be reliable, as precision is of especially high importance in criminology.

Research surrounding the application of linguistic features has been done in several languages, including, for instance, German [9], where the author looked into the usefulness of certain features when it comes to identifying authors. They relied on likelihood ratios (or LR for short) for this purpose. However, in some languages -particularly Hungarian - less research has been done. In Hungarian, feature extraction has mainly been done by hand, and there are tools out there for linguistic analysis, for example, Laurence Anthony's software [2]. Considering how morphologically rich and complex this language is, the

possibilities for automated linguistic feature extraction for forensic applications are vast, and research in this field is both highly motivated and necessary. In this article, we will detail our approach for this task aided by the Hungarian linguistic processing tool called magyarlanc, which was developed using machine learning [22], and inspect our system that allows experts to analyze texts through a graphical user interface. We will also be discussing the implications of our research and potential future directions.

2 Background

The basis of criminalistic or forensic linguistics is the unique use of a language (idiolect). An idiolect is a way a person uses a language, their relationship to it or even how they came to learn it. People master a language (in our case, the Hungarian language) to some degree through socializing. The tools it gives us are then varied, filtered and combined in our heads. The way we go about this can depend on our close environment, our social status, our education, what we read, how well we grasped the language, our age, our gender and other factors as well ([19]). Some aspects of how we use a language depend on the topic or the text's genre, or maybe even the target audience, but other aspects can be subconscious decisions, such as repeating words, connectives and function words used, the complexity of the sentences, how often we use multiword expressions and phrasemes and the way we use them, etc. Experts compare the suspect's texts with the incriminating text using these features, then decide how likely it is that the authors match. Features influenced by subconscious decisions can be especially reliable for this goal, as they are much more difficult to fake. Comparisons in our case are primarily made on the morphological, syntactical, semantical and pragmatic levels, both quantitatively and qualitatively. The expert can also compare features of the suspect's text with features of a generalized corpus, then looks for any abnormalities, features that differ from the norm greatly.

Criminals acting in an online environment can, in most cases, be traced, and this approach is usually both faster and cheaper than analyzing texts with linguistic experts. However, this method of identifying criminals does not always work, especially with modern tools that mask the author's identity, such as virtual private networks (VPN) and browsers that enable anonymous communication; for example, The Onion Router [3]. In such cases, one may have to rely on analyzing the text itself. This does not come without its own difficulties, however, as texts of this nature are often very differently written than texts from the real world, such as ones people would send in a mail. When online, people usually write very casually, not paying much attention to punctuation or correct spelling. Slang is often used alongside emojis. Abbreviations are very common, affixes are somewhat ignored, words are often repeated, and sometimes even the language used is mixed: seeing IM (Instant Messaging) texts that are written in half English and half Hungarian is increasingly common. Even in these cases, identifying the author behind the anonymous messages can be of critical importance.

In many countries forensic linguistics is an active field of research [4,11,17,13,15], and much work has been done regarding computer analysis for forensic purposes as well, with a wide range of bibliography [5,8,14,16,18,21], even examining cross-language methods [7,10]. In Hungary, however, the area is a lot less covered, with only six linguistic experts even being registered in The Ministry of Justice, and the only forensic institute working with linguistic experts is the Institute for Expert Services within the Special Service for National Security. It is no coincidence that research involving authorship analysis using computers is also done within the organization. Experiences of foreign partners and international research also help direct their own research in order to deliver to clients as efficiently as possible.

The organization's modernization in this field is moving in the following directions:

1. Computer text analysis: The first stage is to develop text analysis using computers on as many linguistic levels as possible. For texts that can be processed either manually or with existing software, this should be automated. Depending on the experiences, this could be followed by an attempt to implement feature extraction with computers from shorter or incomplete sentences as well.
2. Computer text comparison: Automatic one-to-one text comparisons could be implemented with an expert's control, utilizing computer analysis that reliably and effectively processes text on multiple linguistic levels. The expert could prioritize certain linguistic features by determining their value in distinguishing authors. The goal is to automate this as well, after some testing and fine-tuning.
3. Computer text comparison on an existing database: Point number two could serve as a basis for one-to-many text comparisons, which can then be used to compare newly written anonymous texts with texts from previous cases in order to find patterns and schemes (for example, one could infer some general traits of threat letters). It could also be used for serial offenders (the same person threatening/blackmailing/slandering other people, at different times, in different places) to identify reoccurring motifs. The goal is for a computer to make rankings of texts based on how similar they are (according to some kind of score).
4. LR-based ranking: Likelihood Ratio based identification using a population database (referential corpus). The computer compares the text in question both with a sample text and with the database, then outputs an LR value indicating how much more likely it is that the text is authored by the suspect instead of anyone else. This method is largely independent of the human expert. The computer determines whether the suspect is guilty or innocent. The introduction of this method to the field of expert linguistics is not underway yet; the composition of the populational database and the effectiveness and reliability of feature extraction are critical for this process.

The first stage of our current research involves developing computer analysis and comparisons, alongside making feature extraction require less and less

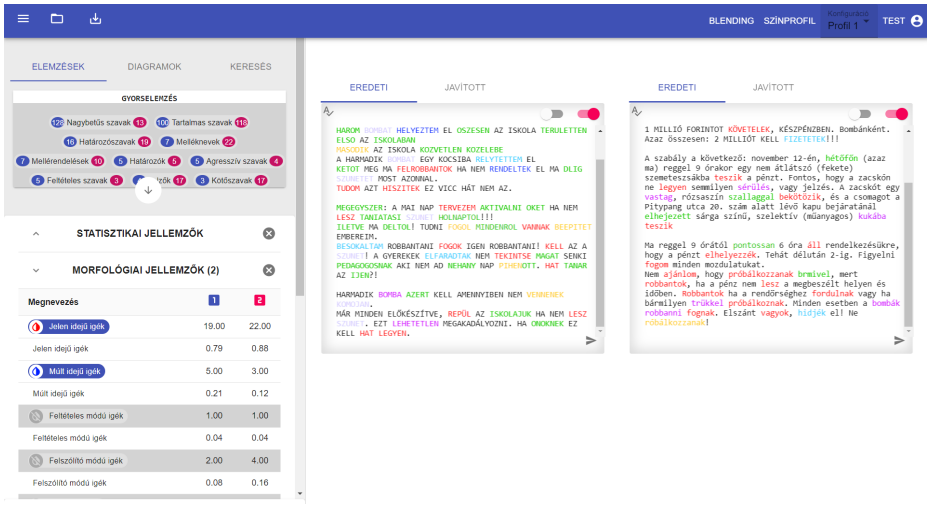

Fig. 1: A preview of the system in use.

human input. The extracted data should be statistically comparable in order to enable the computer to make a judgment about the likelihood of texts being written by the same authors, relying solely on statistical data.

3 Methods

In this section, we will be inspecting how our system operates and the possibilities it opens up to linguistic experts. We will demonstrate the system using example texts about bomb threats in Hungarian. We also present data on a real corpus on which experiences are to be conducted in the next section.

3.1 Automatized Extraction

As seen in Figure 1, our system accepts one or two texts and then extracts their numerous linguistic features automatically. The features are categorized into nine different categories: statistical, morphological, part-of-speech, semantic (for example, aggressive words), semantic uncertainty (for example, peacock words), semantic emotions (for example, words used to express anger), syntactic and indices for textual structure, pragmatic, and finally spelling associated features. Next to each category's name, one can see how many features are currently selected, and one can deselect all of them at once with the X button.

The user can see the quantity associated with each feature for both texts side-by-side, making it simple to compare them. Where it is reasonable to show both a ratio and a quantity for a feature, both are shown as two separate features with the same name. However, the quantity features are easily distinguishable

Fig. 2: A closeup of features in a category, with two of them highlighted.

by their darker grey background. Clicking on features with a darker background highlights the related tokens in the text: for example, by clicking on the past tense verbs feature, one can highlight all the past tense verbs in both texts. The system is made for Hungarian texts (to be used by the Hungarian Special Service for National Security), so the feature names are also Hungarian. Figure 2 demonstrates a possible setting. "Jelen idejű igék" means present tense verbs, while "Múlt idejű igék" means past tense verbs, and these are among the highlighted words. Negative words and misspelled words are also highlighted, the former with purple and the latter with green. The color which the user wants to highlight words can be freely adjusted, and they can also be highlighted by making them bold, italic or underlined. The configuration window can be seen in Figure 3 If a word is highlighted multiple times (for example, if the word is a verb and also negative, and both verbs and negative words are highlighted), the colors are combined by default, but this behavior can also be configured.

Most of these features are calculated using magyarlanc [22], a Hungarian text analysis tool developed through machine learning. More simple features, such as dates (for example, the name of a month, or 2022-04-21) or racist words, are detected simply by using regular expressions or customizable word lists. The system can also detect location names, names of people or names of organizations with the aid of a Named-entity recognition (NER), using the huBERT [12] model. huBERT is a machine learning model based on the groundbreaking BERT

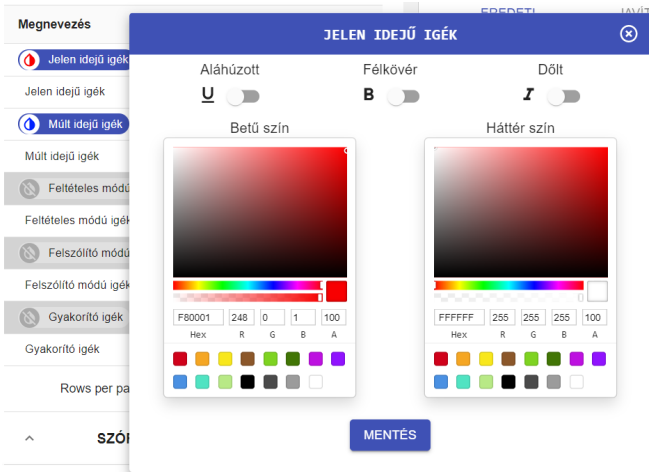

Fig. 3: The highlight customization window of a feature.

(Bidirectional Encoder Representations from Transformers [6]). Unlike the regular BERT model, huBERT was trained specifically for the Hungarian language, on Hungarian corpora, using the structure and mechanics of BERT.

For texts that have very long sentences, text analysis with magyarlanc can occasionally take a long time, so there is an option to use fast-analysis on the text (this can be toggled in the upper-right corner of the text box with the left-side button). The fast analysis automatically puts punctuation marks after a certain number of tokens in the sentence if it is too long. This process is repeated until the sentence is deemed short enough for a fast analysis. Normally, the text is analyzed upon changing the contents of the text box and waiting for a few seconds, but this behavior can be disabled in favor of manually starting the analysis by using the right-side toggle in the upper-right corner of the text box.

For detecting misspelled words, we used HunSpell, a popular free spell checker and morphological analyzer library [1] with some rule-based distintions. An example of detecting misspelled words can be seen in Figure 4. The figure shows multiple misspelled words, "‘hétőfőn"’ is supposed to be "‘hétfőn"’, "‘szallaggal"’ is spelled with only one l ("‘szalaggal"’), "‘elhejezett"’ is written with the wrong j/ly character (correctly, it is "elhelyezett"). In the Hungarian language, the letter "j" is sometimes written with the letter "ly" instead: it depends on the word which letter should be used. This can lead to spelling errors very often, so there is even a separate feature for this in our system, called "j-ly hibák" (j-ly errors). If the linguistic expert analyzing the text wishes to see the text written in the correct form, they can ask the computer to make a corrected version of the text, by pressing the button in the top left corner of the text box. Afterwards, the correct version of the text will appear in the separate "Javított" (corrected) tab.

Fig. 4: A closeup of an analyzed (left) and corrected (right) text. Misspelled words are highlighted with green. All the previously misspelled words are now spelled correctly.

The corrected text is also made using Hunspell. It is important to have the option to correct texts automatically because some features will be more accurately evaluated in this form of the text. For example, misspelled verbs will not be recognized as verbs (as most of the time, the misspelled word is not even a proper, existing word in the language). For this reason, text analysis (except for spell checking, naturally) is automatically done on the corrected form of the text instead of the original form. The expert may also wish to see the corrected form of the text, so it is important to let them see that as well. This can be especially important for texts that have a great number of misspelled words. They might also want to correct the text (or some parts of the text) manually, instead of letting the computer handle it, so text in the corrected tab can be manually edited as well, and the automatic correction only shows in the text box after pressing the corresponding button.

In order to make the expert's work faster and smoother, there are several helper functions available in the system. The most significant differences are displayed in a quick preview as seen on the left side of Figure 5. While using the system, the expert is also logged in with a profile that they can use to save highlight and word list configurations.

On the diagrams tab, visualized analysis of the data is also available, this is visible on the right side of Figure 5.

Users can also search in the text in the search tab. While searching, they can use filters that examine whether certain words or lemmas of a specific POS type are within a given range. For example, if the text contains "three bombs, two bombs", and one searches for the "bomb" lemma, normally, they would find both occurrences. By using a filter such as "the word three must also appear before the word within a range of 1 token", only the first occurrence will be counted. An example is displayed in Figure 6

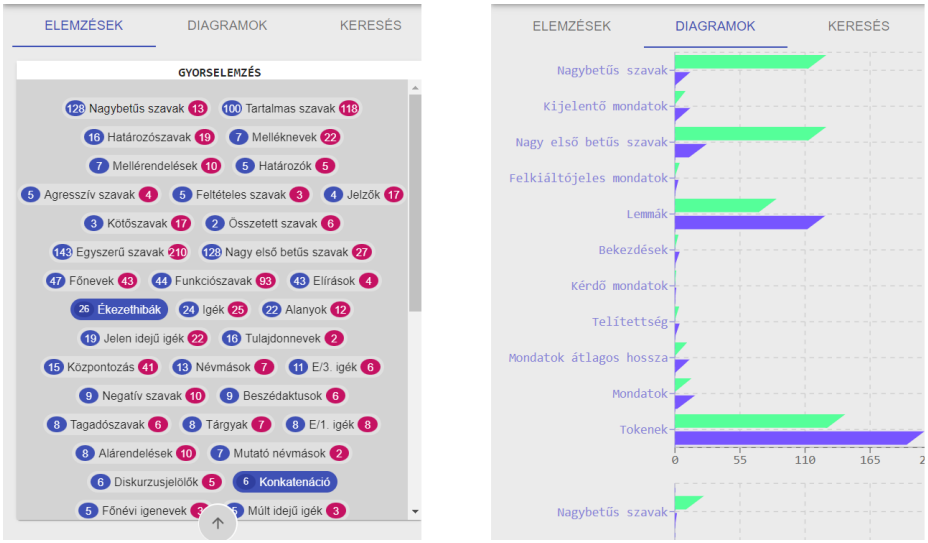

Fig. 5: Left: A quick overview of the analysis as seen in the system. Right: Diagrams of the system. Diagrams alternate between feature values and ratios (first diagram is for values, second is for ratios). Green bars are for the text in the left box, while blue bars are for the text on the right. Exact values can be seen by hovering the cursor over the bars.

3.2 Experimental Corpus

During our experiments, we will examine how various linguistic features can help determine whether two corpora were authored by the same person and also how they can give hints about the author's identity. The texts themselves are written in the Hungarian language, but we provide a brief summary of their contents in English:

Corpus A has texts from postcards and letters written by a woman with a high school diploma as her highest form of education. She is jealous of her former partner's current relationship and strives to ruin it by sending anonymous messages regarding an affair to the man and his current partner. **Corpus B** contains mails sent by a hunter with a university degree. The man falsely accuses his hunting mates of illegally obtaining highly expensive deer horns. **Corpus C** contains letters and e-mails sent by a man. In these texts, he is blackmailing a bank's CEO, attempting to extort money, and threatening to expose their supposed tax fraud and misappropriation of public funds. His highest form of education is a university degree in economics. The last corpus, **corpus D**, contains postcards, letters and e-mails written by a woman with a university degree in liberal arts. She is jealous like author A and is in a similar situation. She, however, goes even further, going as far as sending defamatory texts to the new girlfriend's workplace and even sending bomb threats. She also uses many aliases, using both fictional

and real (other people's) identities. She also sent anonymous letters and used a great number of different signatures, such as "a father", "a benefactor' or "a grandmother". The ages of all four authors are between 33 and 55 (middle-aged), and they all contain both handwritten and digital letters, except for corpus C, which doesn't contain handwritten texts. Further information about the corpus can be found in Table 1.

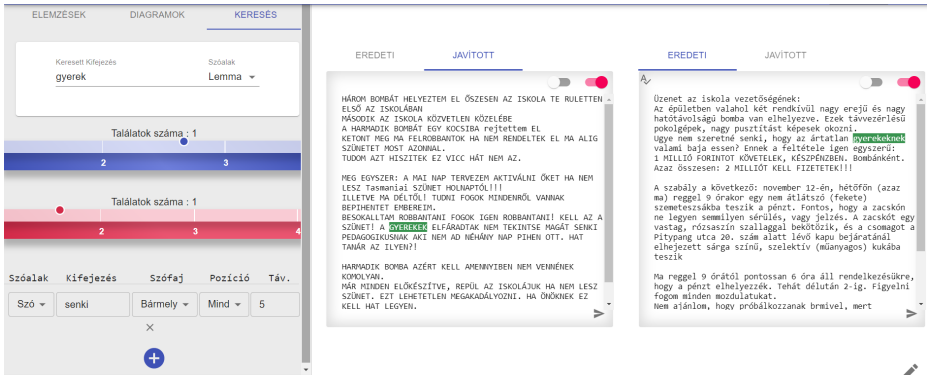

Fig. 6: An example of the advanced search. The lemma "gyerek" (child) is searched, and the word "senki" (nobody) has to appear in a range of 5 words in either direction. In this case, the POS type of the "senki" word is not considered. At the blue and red bars, the dot visualizes the location of the word in the text.

Table 1. Parameters of the corpus used in the experiments

Author	Number of tokens	Number of sentences	Number of documents
A	7 101	528	10
B	2 932	97	6
C	5 940	296	13
D	15 254	812	32
Total	31 227	1 733	61

4 Results

This section displays the most important results of the comparative analysis made between the texts of the four authors detailed above. Only the ratios of the features will be considered, as the available text between authors varies greatly, and the quantities of the linguistic features would reflect this greatly.

4.1 Statistical Features

Table 2 shows the quantified results of the most important statistical features. Note that the words length rate is calculated as follows: (total length of tokens, except for punctuations) / (total number of tokens - total number of punctuations)

Table 2. The most important statistical features.

Feature	A	B	C	D
Rate of lemmas	0.17	0.22	0.18	0.12
Saturation	1.06	2.45	1.58	0.84
Words length rate	5.39	7.02	6.38	5.77
Rate of fully uppercase words	0.04	0.03	0.05	0.06
Rate of capitalized words	0.08	0.10	0.09	0.10
Average sentence length	13.45	30.23	20.07	18.79
Rate of declarative sentences	0.61	0.55	0.75	0.93
Rate of sentences ending with an exclamation mark	0.33	0.20	0.12	0.05
Rate of interrogative sentences	0.06	0.26	0.14	0.02

4.2 Morphologic features

Table 3 shows the quantified results of the most important morphologic features.

Table 3. The most important morphologic features. Sg-singular, Pl-plural.

Feature	A	B	C	D
Rate of present tense verbs	0.67	0.50	0.53	0.63
Rate of past tense verbs	0.34	0.49	0.48	0.38
Rate of conditional verbs	0.08	0.06	0.03	0.07
Rate of imperative verbs	0.07	0.05	0.08	0.09
Rate of frequentative verbs	0.13	0.11	0.17	0.17
Rate of causative verbs	0.00	0.02	0.01	0.01
Rate of modal verbs	0.02	0.02	0.07	0.03
Rate of 1Sg verbs	0.35	0.30	0.18	0.31
Rate of 2Sg verbs	0.10	0.00	0.01	0.09
Rate of 3Sg verbs	0.46	0.48	0.57	0.49
Rate of 1Pl verbs	0.04	0.00	0.07	0.03
Rate of 2Pl verbs	0.00	0.00	0.00	0.00
Rate of 3Pl verbs	0.05	0.22	0.17	0.09
Rate of superlatives	0.02	0.02	0.02	0.00
Rate of comparatives	0.05	0.02	0.03	0.04
Rate of plural nouns	0.08	0.08	0.12	0.14

4.3 Part-of-speech features

Table 4 shows the quantified results of the most important part-of-speech features.

Table 4. The most important part-of-speech features.

Feature	A	B	C	D
Rate of verbs	0.14	0.07	0.09	0.13
Rate of nouns	0.20	0.30	0.26	0.19
Rate of adjectives	0.06	0.07	0.07	0.06
Rate of numerals	0.01	0.01	0.01	0.01
Rate of adverbs	0.12	0.05	0.08	0.12
Rate of conjunctions	0.08	0.05	0.05	0.08
Rate of pronouns	0.08	0.05	0.04	0.11
Rate of proper nouns	0.00	0.04	0.02	0.00

4.4 Syntactic features

Table 5 shows the quantified results of the most important syntactic features.

Note: the rate of clauses is calculated as follows: total number of clauses / total number of sentences

Table 5. The most important syntactic features.

Feature	A	B	C	D
Rate of subjects	0.05	0.03	0.06	0.06
Rate of objects	0.05	0.04	0.04	0.04
Rate of adverbials	0.04	0.02	0.02	0.04
Rate of attributives	0.05	0.08	0.07	0.05
Rate of coordinations	0.05	0.08	0.08	0.06
Rate of subordinations	0.05	0.03	0.03	0.05
Rate of clauses	2.12	2.67	2.31	2.64
Rate of simplex sentences	0.39	0.29	0.34	0.32
Rate of complex sentences	0.61	0.71	0.66	0.68
Rate of clauses in complex sentences	2.84	3.35	2.99	3.40
Rate of complex sentences with two clauses	0.30	0.25	0.30	0.23
Rate of complex sentences with three clauses	0.20	0.22	0.19	0.20
Rate of complex sentences with four clauses	0.05	0.14	0.10	0.14

4.5 Semantic features

Table 6 shows the quantified results of the most important semantic features.

Table 6. The most important semantic features.

Feature	A	B	C	D
Rate of positive words	0.02	0.02	0.03	0.02
Rate of negative words	0.02	0.02	0.03	0.02
Rate of simplex words	1.00	0.98	0.99	0.99
Rate of compound words	0.02	0.04	0.03	0.02
Rate of content words	0.54	0.57	0.53	0.54
Rate of function words	0.46	0.43	0.47	0.46
Dividedness of sentence units	1.53	2.95	2.33	1.79
Saturation rate for sentence units	3.40	6.44	4.62	3.82
Contentness	7.22	17.20	10.69	10.08

4.6 Pragmatic features

Table 7 shows the quantified results of the most important pragmatic features.

Table 7. The most important pragmatic features.

Feature	A	B	C	D
Rate of private verbs	0.16	0.08	0.13	0.16
Rate of public verbs	0.07	0.06	0.07	0.06
Rate of speech acts	0.20	0.15	0.14	0.19
Rate of punctuation	0.19	0.23	0.25	0.18

5 Discussion

Overall, between authors A, B, C and D, intuitively, authors B and C should be somewhat close as both their gender and education match, and the nature of their letters are also somewhat similar. Authors A and D are also very similar, as both are women and both are in a very similar situation: however, their degree of education differs, and their methods differ greatly as well. Generally, A and D -although similar- are not as close to each other as B and C, and looking at it as a whole, author A seems to be an odd one out in general, being the only one with only a high school diploma. We will now investigate how the results reflect this.

There are a number of interesting distinctions one can make between the four corpora by analyzing the results. Author B, for instance, seems to prefer writing long and complex sentences, which are also quite saturated. Not only are his sentences lengthy, but they also contain many meaningful words. This is all contrary to author A, who uses the shortest sentences amongst the four, and said sentences are also one of the least saturated: though D's sentences are a little less saturated even. The average length of the words used is also

highest in the case of B and lowest in the case of A. The same goes for the rate of sentences with four clauses, which can be considered very complex. On the other hand, author A writes a lot of sentences with two clauses, she's only matched in this regard by author C, so one can't say that A avoids sentences with multiple clauses altogether, but she does seem to feel comfortable with less than four clauses. The rate of clauses shows that altogether, author A avoids complex sentences the most, and author B prefers them the most, although he's very closely matched with D. This may point to author D having a degree in liberal arts, which can also indicate a humanistic sense.

One should also note the different types of sentences the authors use. D appears to use declarative sentences the most. In fact, the vast majority of her sentences are declarative. One can also infer that B uses a lot of interrogative sentences, so it appears they tend to ask a lot of questions. As for exclamatory (and imperative and conditional) sentences, it appears that A prefers using them the most. It's an interesting contrast between her and D, as both are in a similar situation, yet D avoids them the most. Perhaps D attempts to mask her emotions more than A. She also uses a lot of aliases. It is apparent that she does try to cover her tracks.

One can also detect differences between the four by the way they use verb tenses. A and D both talk a lot more about the present than both B and C. The past tense appears to be preferred by B and C, compared to A and D. This is arguably somewhat of a surprise, as A and D would intuitively have a lot to reflect on, as they are both trying to revive a past relationship. However, neither seems to behave accordingly and instead, they appear to focus on the now in their texts. As for B and C, the most probable reason they both prefer the past tense is that they are both accusing another party of having done something, so they talk a lot about what another person has done in the past. Looking at it this way, it makes sense that A and D focus on the present, as in order to break up their ex's current relationship, they would probably prefer to point out flaws in their new relationship rather than the old one, and flaws in the people they are in the present. The usual verb tenses of one's text can indicate an underlying personal preference, but as it is visible, it is more likely to be influenced by the subject and goals of the text than stylistic choices.

Another point of interest is the person each author focuses on in their texts. A, B and D all seem to be focusing a lot on themselves, but C sticks out in this regard: he focuses on himself a lot less, only barely using first-person conjugated verbs compared to the other three. The rate of third-person verbs appears to indicate that all four focus on a third party a lot, especially C, but it's worth noting that in the Hungarian language, in formal writing, the second-person verb form is often discarded in favor of the third-person form. C actually does write in this manner, which is the true reason his rate of third-person verbs is this high. A B and D, however, all talk about a third person in their messages. Taking all this into consideration, it's evident that one must be very careful with interpreting features, as their meaning can change very significantly from language to language.

We can also use the system's search function in order to uncover some very interesting findings: A has the tendency to misuse the Hungarian connective "is" by incorrectly concatenating it to the preceding word. In Hungarian, the "is" connective should always be written separately from the preceding word. The other three authors do not make this same mistake. Only B uses the "összefüggésben" (in Hungarian) structure. We say something is in "összefüggésben" with something else if something is related to that something else, but there are other ways of expressing this too. D has some unique traits as well, she is the only one who ends words with the "képpen" affix, and she uses it four times, showing a clear tendency. In Hungarian, the "képpen" affix is easily replaceable, so D did not necessarily have to use it in any of the instances. She also often uses the word "kép" (picture in English) in her structures, for example, "out of the picture". This structure is common in both the Hungarian and English languages. This, too, can easily be phrased differently, and the other three authors do not use it as often. There is also an example of a word that two authors, B and D, both use, but even though they use it to express the same thing, they use it in different ways: In Hungarian, the word "továbbá" (in English, this translates to "furthermore", or "in addition") can be used both at the start of a sentence or somewhere in the middle. B consistently uses it mid-sentence, while D consistently uses it at the beginning of the sentence.

6 Conclusions

The goal of this research was to show that through the aid of machine learning, helping linguistic experts in identifying authors even behind short, digitally written texts is possible. We implemented a highly customizable system that combines machine learning technologies and simpler technologies to allow automated feature extraction from Hungarian texts. The system allows the expert to smoothly analyze and visualize the data with numerous tools, such as highlighting certain features, correcting texts and advanced searching in texts. We discussed how the experts can use the data and these tools to get hints about the author and ultimately identify them by pairing their texts with texts from known authors. In the future, our goal is to also automate the process of determining whether two bodies of text (or a query text and a corpus) were authored by the same person based on the data extracted and could even contribute to comparison to a population database to inspect traits of criminals writing malicious texts.

Acknowledgements The research (conducted by the Special Service for National Security and the University of Szeged) was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004).

References

1. Hunspell, http://hunspell.github.io/, accessed: 21-03-2022
2. Laurence anthony's linguistic analysis software, https://www.laurenceanthony.net/software.html, accessed: 21-03-2022
3. Tor project, https://www.torproject.org/, accessed: 21-03-2022
4. Coulthard, M., Johnson, A.: The routledge handbook of forensic linguistics (2010)
5. 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)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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. pp. 4171–4186. Association for Computational Linguistics (ACL) (oct 2019)
7. 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)
8. 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)
9. Ishihara, S.: Strength of forensic text comparison evidence from stylometric features: a multivariate likelihood ratio-based analysis. The International Journal of Speech, Language and the Law **24**, 67–98 (2017)
10. 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)
11. McMenamin, G.: Advances is forensic stylistics. CRC Press (2002)
12. Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)
13. Nini, A.: Authorship profiling in a forensic context. In: Ph.D. thesis (2014)
14. Nirkhi, S., Dharaskar, R., Thakare, V.: Authorship verification of online messages for forensic investigation. In: Procedia Computer Science. vol. 78, p. 640–645 (2016)
15. Olsson, J.: Forensic linguistics: An introduction to language, crime, and the law. Bloomsbury Publishing, New York (2004)
16. Rexha, A., Kröll, M., Ziak, H., Kern, R.: Authorship identification of documents with high content similarity. In: Scientometrics. vol. 115, p. 223–237 (2018)
17. Shuy, R.: Linguistics in the courtroom: A practical guide. Oxford University Press, New York (2006)
18. Sousa-Silva, R.: Computational forensic linguistics: An overview of computational applications in forensic contexts **5**, 118–143 (12 2018)
19. Szilák, J.: Az írásszokások néhány formai jegyének hátteréről. In: Belügyi Szemle. vol. 16, p. 67–68 (1980)
20. Veronika, V., András, K., Eszter, F., László, V.: A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei. XVII. Magyar Számítógépes Nyelvészeti Konferencia p. 275–288 (2021)
21. Zhang, C., Wu, X., Niu, Z., Ding, W.: Authorship identification from unstructured texts. In: Knowledge-Based Systems. vol. 66, p. 99–111 (2014)
22. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of hungarian. In: Proceedings of RANLP. p. 763–771 (2013)