Zhytomyr Ivan Franko State University Journal.
Philological Sciences. Vol. 1 (104)

Вісник Житомирського державного
університету імені Івана Франка.
Філологічні науки. Вип. 1 (104)

ISSN (Print): 2663-7642
ISSN (Online): 2707-4463

УДК 811.111' 367.7 : 659.3
DOI 10.35433/philology.1 (104).2025.126-133

POTENTIAL OF AUTHORSHIP VERIFICATION THROUGH AUTOMATIC SYNTACTIC ANALYSIS

O. V. Hyryn[*]

This paper discusses the potential for using Automatic Syntactic Analysis to verify authorship, particularly in educational settings. The focus is on the creation of a tool with the project name AVASA to detect an individual's unique syntactic signature – a "linguistic fingerprint" shaped by their sentence structure, word choice, and punctuation use. The goal is to ensure that students submit original work, independent of AI-generated content or plagiarism, while also tracking their evolving writing style.

The paper outlines existing authorship attribution tools, which analyze features like sentence length, clause structure, and function word usage to distinguish between authors. However, these tools typically focus on static profiles for forensic or cultural purposes, while the suggested tool is intended to handle the dynamic nature of student writing as it evolves throughout their education.

Key steps in the authorship verification tool include data collection, feature extraction, and the creation of a measurable syntactic profile. The system begins by gathering a diverse set of text samples from the student, which are analyzed to extract features such as sentence length, complexity, verb tense, and punctuation patterns. These features are quantified and compared with future submissions to update the student's profile dynamically. If a new submission significantly deviates from the profile, it is flagged as suspicious, potentially indicating plagiarism, AI use, or ghostwriting.

The system's dynamic nature ensures that it adapts to students' developing writing skills over time, while providing teachers with objective data on linguistic progress. AVASA also is intended to support multi-author projects by segmenting documents and attributing sections to individual contributors based on their syntactic profiles. Challenges include ensuring the first submission is accurate, defining thresholds for syntactic deviations, and handling diverse writing styles. Despite these challenges, AVASA holds promise for improving authorship verification in educational and non-educational settings.

Key words: *automatic syntactic analysis, syntactic signature, authorship verification, syntax dynamics.*

[*] Candidate of Philological Science (PhD), Associate Professor
(Zhytomyr Ivan Franko State University)
hyryn-o@zu.edu.ua
ORCID: 0000-0002-3641-2440

Zhytomyr Ivan Franko State University Journal. Philological Sciences. Vol. 1 (104)
Вісник Житомирського державного університету імені Івана Франка. Філологічні науки. Вип. 1 (104)

ПОТЕНЦІАЛ ПЕРЕВІРКИ АВТОРСТВА ТЕКСТУ ЗА ДОПОМОГОЮ АВТОМАТИЧНОГО СИНТАКСИЧНОГО АНАЛІЗУ

Гирин О. В.

У статті описано потенціал застосування автоматичного синтаксичного аналізу для перевірки авторства текстів, зокрема в освітніх установах. Основну увагу зосереджено на створенні інструменту під проєктоною назвою AVASA для виявлення унікального синтаксичного підпису людини – "мовного відбитка", сформованого структурою речень, вибором слів і використанням пунктуації. Мета полягає в тому, щоб переконатися, що студенти подають оригінальні роботи, незалежно від вмісту, створеного штучним інтелектом, або плагіату, а також відстежувати динаміку їх синтаксичного стилю.

У статті описано наявні інструменти визначення авторства, які аналізують такі характеристики, як довжина речень, структура пропозиції та використання службових слів, щоб визначати авторів. Однак ці інструменти зазвичай зосереджені на статичних профілях для криміналістичних або культурних цілей, тоді як запропонований інструмент призначений для обробки динамічного характеру письмового тексту студентів, який розвивається протягом їхнього навчання.

Ключові кроки в інструменті перевірки авторства вміщують збір даних, виділення ознак і створення вимірного синтаксичного профілю. Система починає зі збору різноманітних зразків тексту від учня, які аналізують, щоб виділити такі характеристики, як довжина речення, складність, час дієслова та шаблони пунктуації. Ці функції кількісно оцінюють та порівнюють із майбутніми матеріалами для динамічного оновлення профілю студента. Якщо новий текст значно відрізняється від профілю, він позначається як підозрілий, потенційно вказуючи на плагіат, використання штучного інтелекту або написання із залученням сторонньої допомоги.

Динамічний характер системи гарантує, що вона адаптується до розвитку навичок письма в учнів із часом, надаючи вчителям об'єктивні дані про мовний прогрес. AVASA також передбачає підтримку багатоавторських проєктів шляхом сегментації документів і приписування розділів окремим учасникам на основі їхніх синтаксичних профілів. Проблеми містять забезпечення точності першого подання, визначення порогових значень для синтаксичних відхилень та обробку різноманітних стилів письма. Незважаючи на ці проблеми, AVASA обіцяє покращити перевірку авторства в освітніх і ненавчальних умовах.

Ключові слова: *автоматичний синтаксичний аналіз, синтаксичний підпис, верифікація авторства, динаміка синтаксису.*

Defining the problem. In our previous research on natural language processing (NLP) [2] authorship verification was mentioned as one of possible tasks for automatic syntactic analysis.

Authorship verification is a research subject in the field of digital text forensics that concerns itself with the question, whether two documents have been written by the same person [1]. It is also aimed at reliable automatic authorship detection for an anonymous piece of text (for example, in forensic linguistics). Namely, automatic verification of both vocabulary and especially syntax can conclude with high probability whether an anonymous piece of text and a known text have been written by the same author. Whereas a similar task, text authenticity verification enables defining whether a piece of text was written by a specific author.

In this paper we substantiate the principles and perspectives of using automatic syntactic analysis (ASA) for the attribution of authorship for students' submissions based on their dynamic syntactic profile.

Analysis of Previous research. The idea that every individual has a unique syntactic signature, a linguistic fingerprint [4] based on their habitual use of sentence structure, word choice, punctuation, and syntactic patterns, is an evolving hypothesis [3; 6]. This concept lies at the intersection of linguistics, computational modeling, and stylistic analysis. The challenge, however, is to develop an effective tool that can accurately and dynamically track an individual's evolving syntactic style, detect plagiarism, and even identify the use of AI in text submissions [4]. This idea has the potential to transform how we approach authorship attribution, plagiarism detection, and the identification of AI-generated content, particularly in educational settings.

Authorship attribution has long been a critical area of study in computational linguistics and forensic linguistics. Several tools and approaches have been developed to quantify and identify the unique writing style of an author. These tools employ a variety of linguistic features, including syntactic, lexical, and stylistic markers. Among the existing tools and approaches we can name the following: JStylo: an open-source software platform, that is widely used in forensic linguistics to analyze writing styles using features such as word frequency, punctuation, and syntactic structures; Signature Stylistics: a program that examines both vocabulary and syntax to identify an author's distinctive style; The Authors' Styles Project investigates how function word frequency, sentence length, and other syntactic features can be employed to attribute authorship to disputed texts; Stylistic Profiling: the method, that analyzes an author's use of syntax, punctuation, word choice, and sentence structures to build a profile.

These tools have successfully demonstrated that syntactic markers such as sentence length, sentence complexity, and clause structure can distinguish authors, providing a basis for automated authorship attribution. They however are aimed at resolving individual forensic and cultural tasks with the build-up of a linguistic profile for a specific author and its subsequent comparison with the new or a disputable text. In education, where academic integrity is the basic principle, authorship needs to be verified on a daily, even hourly basis. Moreover, beside proving that a certain work was created by a certain author, the tool needs to account for the possible change in the student' writing style as it is one of the studying goals at a linguistic school.

Existing anti-plagiarism tools compare a submitted text with the existing database. AI-generated text poses a unique challenge in plagiarism detection as they are recognized as unique by the traditional anti-plagiarism tools. However if the bottom line for the authorship verification is the author themself, the authorship attribution of an AI generated or automatically translated text will be impossible. AI-generated texts often exhibit consistent, "non-human" syntactic features: overuse of simple sentence structures, unusual patterns of punctuation – AI text might use punctuation differently, often overusing commas or periods in unnatural places, repetitive syntax: AI models often repeat certain sentence structures, making them distinguishable from human-authored text.

The need for a new approach in verifying a student's authorship in academic writing is the key motive for creating a tool, where it will be impossible for a student to submit any other but independently formulated written text.

Results and Discussion. The first critical step in creating a tool for authorship verification (project name AVASA: Authorship Verification through Automatic Syntactic Analysis) to detect an author's syntactic signature is the development of a comprehensive, quantifiable profile of the individual's unique writing style. This "syntactic fingerprint" is normally constructed by analyzing a sufficiently large and diverse sample of text produced by the target author.

By capturing their stylistic choices across different types of text and contexts, a more accurate and reliable profile can be established. This process can be broken down into several key stages:

1. Data Collection. The foundation of any accurate syntactic signature begins with a reliable data collection. A large variety of text samples from the target author should be gathered to ensure the

analysis reflects their natural writing style in different contexts. The more diverse the corpus, the better the resulting profile will capture the full spectrum of the author's writing. For instance, for students these samples could include essays, texts of their speeches, articles, reports. By combining text from various genres and medias, a more holistic view of the author's syntactic style is achieved, which ensures that the tool accounts for their writing idiosyncrasies in both formal and informal settings. However, in our case with students the profile will start with the first submitted written assignment, which could be written at the first corresponding class at the educational establishment. Every next entry will update the student's profile should the student's authorship be verified by the tool. At the initial stage it is essential that the first text be 100% written by students themselves, without the use of AI, automatic translation, or copied parts from other sources. Therefore the students need to be advised that it is in their interests to complete the first assignment diligently and independently. In failure to do so, all their further written submissions, even those, indeed written by them would be recognized as those with unattributed authorship. Therefore it might be a good idea not to academically evaluate the initial assignment, so that the students wouldn't be motivated to show results that would be better than they are really capable of.

2. Feature Extraction. Once the data collection phase is complete, the next step is feature extraction, where the raw text is processed to derive measurable syntactic features. These features are key to identifying an author's unique writing style, and they include a wide range of syntactic and structural elements, such as:

- Average sentence length: examining the average number of words per sentence and how sentence length varies across the text. This helps identify writing styles that prefer laconic sentences or more complex, lengthy structures.

- Average clause length: measuring the length of clauses within sentences to assess syntactic complexity.

- Proportion of simple vs. complex sentences and short vs long sentences identifying the ratio of simple sentences to compound and complex ones, which reflects the author's style of combining various types of syntactic structures, preference for simplicity or complexity.

- Use of specific syntactic structures: identifying patterns such as the use of *there is/there are* constructions, modal constructions, relative clauses, appositions, or other syntactic constructions that may be unique to an individual author.

- Use of function words: analyzing the frequency of function words (e.g., "*the*," "*is*," "*of*") provides insight into an author's style. Function words tend to be highly individual and are often used consistently across different genres [3].

- Proportion of passive vs. active constructions: insight into the balance between active and passive voice provides additional clues to the syntactic signature, with some authors exhibiting a preference for one over the other.

- Verb tense and aspect: analyzing the choice of verb tense (present, past) and aspect (progressive, perfect) to see how an author typically conveys time and action.

- Punctuation usage patterns: identifying unique punctuation practices, such as frequent use of commas, semicolons, or ellipses, can be an important marker of an author's style.

- The use of pronouns: the choice of pronouns (first person, third person, singular vs. plural) can be a distinctive syntactic feature, often linked to the author's tone and voice.

-Syntactic tree structures: the most complex feature to be analyzed, but it can provide an ultimate syntactic signature of an author as language users do prefer certain syntactic patterns over other.

3. *Quantification and Representation.* The next step in the syntactic analysis

algorithm is to translate these syntactic features into measurable metrics that can be analyzed and compared. Each feature should be quantified in a way that will allow meaningful comparison between different texts. For example: average sentence length could be defined by the average number of words or syllables per sentence. For a written text the number of words would be a more preferable criterion than the number of syllables, as the number of syllables (as an option, only stressed ones) could be a criterion for oral texts. The latter is not our objective in this study so we sill focus on the number of words in a sentence as a universal feature. For example, the average number of words in a sentence in this paragraph is 21,7. Whereas the rest of the article above shows that the average number of words in a sentence as 23,1. Thus the deviation of 6% within a text is well within the acceptable range, which will have to be empirically defined for different types of texts. For a relatively stylistically uniform text like a scientific article, the deviation cannot be over 10% for the same text and 20% for a dynamic system, that will take into account the capacity of the author to change their style over time.

4. Prioritization. After extracting and quantifying the relevant syntactic features, the next step is to define the key syntactic features, violation of whose deviation range is an immediate flagging of plagiarism. It has to be followed by a group of features, that can deviate, even significantly, but the majority of which still have to correspond to the author's style. And, finally define the features, that contribute to the overall syntactic signature, but can deviate from text to text.

A significant challenge in authorship attribution and plagiarism detection, particularly in educational settings, is the evolving nature of a student's writing style. As students develop their writing skills, their syntax, vocabulary, and overall stylistic choices undergo substantial changes. A static model for analysis may fail to account for these changes, leading to misclassifications and false positives.

The proposed dynamic updating mechanism. That is proposed by this paper is designed to address this challenge by continuously adapting the student's writing profile. It is suggested to implement the following stages of self-improvement:

Syntactic profiling: the system creates an initial profile for a student based on their first text submission. This profile captures key syntactic features listed above. The features are analysed in the way also suggested above and comes up with a set of measurable parametres for the student. It is important that the interface of the tool should have an option of selecting the type of the submitted paper work: an essay, a project work ,a scientific article etc., as each type of work corresponds to a different style of the written speech – from colloquial even informal, to formal and literary. All these styles are obviously characterized by a different syntax, which has to be accounted for during the automatic syntactic analysis and authorship verification.

Updating the profile: with each new submission, the system analyzes the syntax of the text, obtains the new parametres and calculates the difference between the new features and the existing profile. If the new submission falls within an acceptable range (standard deviations), the system updates the profile. We suggest that the old parameter be stored by the system in an archive or a log, personalized for every student. If a submission deviates significantly from the established profile (i.e., beyond the allowable range), it is flagged as suspicious or inauthentic. This anomaly could indicate the use of AI, plagiarism, machine translation or an external author.

Output of the stored data for the analytical purposes.

For students the dynamic feature of the syntactic analysis is crucial as they are not only apt but also are supposed to change while studying. Moreover, the dynamic analysis system will be able to provide a teacher with objective data on how the student's language has evolved over the course of studying or certain period. If the dynamic analysis system saves the history of a certain student's syntactic parametres, these data could also be a source for further analysis. The program could have a function of displaying the student's progress over time in the form of graphs, charts or any other feasible output. The data could help track major milestones in a student's progress like transition from simple to complex sentences, improvement in the use of function words, student's tense and aspect evolution.

Moreover, this type of data could be valuable for scientific research, especially in fields like cognitive linguistics in understanding how language skills develop over time and identifying cognitive patterns in language acquisition (how sentence structure or complexity correlates with cognitive milestones), linguistic methodology where methodologists could use longitudinal syntactic data to define and concentrate upon most effective teaching strategies or develop more effective pedagogical tools for teaching writing. The data could also be used by educational and/or age psychology to study how various factors (e.g., age, education level, or language proficiency) influence syntactic development, how writing style evolves in response to cognitive development or the impact of different factors. Another possibility for research is comparing the development of writing styles across different student groups with the account for age, sex, education, background. The tool could be adapted to track development of writing in different cultural or linguistic contexts, allowing for cross-linguistic studies of writing progress.

These visualizations of the student's dynamics could provide both the student and educators with a clear view of writing progress over time, motivating improvement or identifying areas needing focus, objective feedback on how to improve a student's writing. Insights such as: *"You have been using simpler sentence structures over the last few assignments – try incorporating more complex sentences"* or *"In your writing you do not/excessively use passive voice"* etc.

During the process of education at an educational establishment, there is also team or group work at some projects. In collaborative academic projects, it is common for multiple contributors to have different writing styles. A system that can detect discrepancies between writing styles within the same document would be invaluable in identifying the involvement of AI, ghostwriters, or plagiarism. By breaking the document into segments (e.g., paragraphs, sections), the system can compare each segment against the evolving author profile and flag significant shifts in style.

The interface of the tool will have an option of adding one or more authors to the submitted text and the tool will verify not only the authorship, but also the input by every author, so that their contribution get due evaluation.

Steps for detecting multi-author projects will include several stages. First, the document will be segmented into smaller parts for easier analysis: paragraphs, blocks. Next, syntactic features will be extracted and analyzed for each segment. Then each segment will be compared to the students' dynamic profiles. Segments that match significantly different profiles will be attributed to the certain author. The results of the analysis suggested by the tool could be the following:

Author A – authorship verified for 65% of the text,

Author B – authorship verified for 20% of the text,

Author C – Authorship not found,

Unidentified authorship – 15% of the text.

This conclusion would mean that in case of the claimed 3 authors for the text, the tool concluded that the first author did most of the work, the second author – did less, the third author either did not participate at all or plagiarized their part of the text.

In order to successfully launch the project some of the challenges need to be addressed. Chronologically and logically, the initial profile accuracy: the first submission needs to be representative enough of the student's typical writing style, which may not always be the case, especially in early academic stages. Next the threshold for syntactic deviations has to be empirically defined, which might take at least one academic cycle where the students' syntax will naturally evolve. Setting the appropriate thresholds for detecting significant deviations requires careful tuning to balance sensitivity and specificity. Associated with the latter challenge is diverse writing styles of the same author; some students may have naturally diverse or not traditional writing styles, which could lead to false flags. The system must be refined to accommodate such variability.

Thus AVASA can initially be used in educational setting by teachers, who can use this tool to track students' progress over time while also ensuring that work submitted is genuinely the student's own. It will enable more accurate and personalized assessments of student development.

In non-educational setting it can be used for the purposes of plagiarism and AI detection for the authors with a previous text history. The system will be highly adaptable for detecting both AI-generated content and traditional forms of plagiarism. By continuously updating the author's writing profile, the tool can distinguish between legitimate changes in style and potential plagiarism or external help.

In collaborative work projects where multiple authors may contribute, the tool can identify discrepancies between authors' writing styles flagging instances where external assistance (e.g., ghostwriting, AI use) may have been involved and where claimed authors were included into the list without actual contribution to the project.

Conclusion. The concept of using a dynamic, evolving syntactic signature for detecting authorship, plagiarism, and AI-generated content offers a novel and highly adaptive approach to addressing contemporary challenges in authorship attribution. By continuously updating an individual's writing profile and detecting significant deviations, this system provides a more personalized and accurate way to track a student's progress and ensure academic integrity. Furthermore, its ability to detect multi-author contributions, including those involving AI, makes it a potentially valuable tool for modern educational settings where collaboration and external help are increasingly prevalent.

The new projected AVASA tool will combine multiple innovative features not available in existing systems, making it a new application in both educational and linguistic research contexts: dynamic tracking of a student's syntactic features over time; visualization of writing progress in terms of linguistic complexity, syntactic variation, and other stylistic parameters; providing data for insights into cognitive development, language acquisition, and personal growth in writing, detecting deviations from a student's evolving writing profile to flag AI-generated text or plagiarism, offering data for scientific research on cognitive linguistics, language acquisition, and educational development.

None of the existing tools offers this full spectrum of functionalities, making AVASA a new and innovative application that could be highly valuable in both education and linguistic research.

REFERENCES

1. Halvani, O., Winter, C., Graner L. (2019). Assessing the Applicability of Authorship Verification Methods / *The 14th International Conference on Availability, Reliability and Security (ARES 2019)*. [Online source]. URL: https://arxiv.org/abs/1906.10551 (reference date: 21.01.2025). [in English].

2. Hyryn, O. (2018). Principal Problems of Natural Language Processing Systems. *Studia Philologica*. Iss. 11. Pp. 35–38. [in English].

3. Kestemont, M. (2014). Function Words in Authorship Attribution. From Black Magic to Theory? / *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Pp. 59–66. [in English].

4. Ramnial, H., Panchoo, S., Pudaruth, S. (2016). Authorship Attribution Using Stylometry and Machine Learning Techniques. *Intelligent Systems Technologies and Applications* / S. Berretti et al. (eds.). Vol. 1. Springer International Publishing Switzerland. Pp. 113-125. [in English].

5. Shastry, U. R. (2019). Linguistic Finger-printing in authorship identification. J*ournal of Emerging Technologies and Innovative Research*. Volume 6. Issue 3. Pp. 527–530. [in English].

6. Varela, P, Justino, E., Soares de Oliveira, L. (2011). Selecting syntactic attributes for authorship attribution. *Proceedings of International Joint Conference on Neural Networks. San Jose, California, USA, July 31 – August 5*. Pp. 167–172. [in English].

Стаття надійшла до редколегії: 19.02.2025
Схвалено до друку: 23.04.2025