Towards the Issue of Creating Linguistic Pedagogical Profiles

Linguistic Pedagogical Profiles

Ivan D. Mamaev

Saint Petersburg State University, Russia, [i.mamaev@spbu.ru](mailto:i.mamaev@spbu.ru)

Baltic State Technical University named after D.F. Ustinov, Russia, [mamaev\_id@voenmeh.ru](mailto:mamaev_id@voenmeh.ru)

The study explores the development of linguistic profiles for evaluating teacher effectiveness in pedagogical discourse. It emphasizes the role of language in educational practices and how linguistic features can distinguish more effective teachers from less effective ones. The research utilizes a corpus of transcripts of Russian secondary school lessons that are analyzed using statistical and machine learning techniques. Key linguistic metrics, such as noun-verb correlation, sentence complexity, and lexical density, are used to create profiles of teacher performance. The study found that more effective teachers demonstrate a balanced linguistic way of delivering information, employing varied syntactic structures and diverse vocabulary. In contrast, less effective teachers tend to exhibit repetitive and rigid language patterns, which may impede student comprehension. While the study findings provide insights into the relationship between language use and teaching success, the authors acknowledge limitations in the dataset size and suggest future research to expand on the current findings. The preliminary profiles can serve as diagnostic tools to help in developing better pedagogical strategies for trainee teachers in adopting language use that enhances learning outcomes.

CCS CONCEPTS • Machine Learning • Document Management and Text Processing

**Additional Keywords and Phrases:** Computational Linguistics, Pedagogical Corpus, Stylometry, Correlation

1. Introduction

The study of social communities has long been a central focus of sociological and anthropological research, offering critical insights into the ways individuals interact, share values, and create collective identities. Social communities, whether defined by geographical proximity, shared interests, or cultural practices, play a crucial role in shaping human behavior and social norms. These communities provide a sense of belonging and identity, facilitate the exchange of knowledge, and support the development of social capital.

Among the diverse types of social communities, learning communities stand out as particularly significant due to their profound impact on personal and social development. Learning communities encompass not only the formal settings of schools, colleges, and universities but also the informal networks that emerge within and around these institutions. A learning community can be defined as a network of individuals and groups engaged in the process of teaching and learning, characterized by shared goals, values, and practices that support the development of knowledge and skills [14].

Within this community, pedagogical practices—the methods and strategies used to facilitate learning—play a central role in shaping the educational experience. More effective communication between teachers and students is essential for the successful transmission and assimilation of knowledge, and the language used within these interactions is a critical tool for mediating understanding, fostering critical thinking, and building relationships [24]. Language features, such as morphological, syntactic and discourse patterns, are integral to the functioning of learning communities. For instance, the use of inclusive language can promote a sense of belonging and mutual respect, while the careful selection of instructional language can enhance comprehension and engagement.

The papers collectively demonstrate the power of statistical methods to reveal underlying patterns in language use, whether through syntactic structures, prosodic features, or conceptual metaphors. For analyzing effective teachers' linguistic profiles, these methods are highly relevant. We suggest that the correlation-based insights from these papers can be applied to explore what makes certain teachers' language more effective in educational settings. This work aims to create linguistic profiles of more and less effective teachers in the corpus of Russian pedagogical practices using statistical and machine learning methods. The developed profiles can potentially be used as formal identifiers of lesson success.

1. RELATED WORK

Statistics is significant in linguistic studies as it provides rigorous methods for analyzing linguistic data, enabling researchers to identify patterns, and draw empirically grounded conclusions [19]. It facilitates the quantification of different language phenomena, such as parts of speech, syntactic constructions and stylistic devices. For instance, Baburchenkova [3] delves into the individual style of William Blake through a quantitative analysis of his early lyrics. The main idea of the paper is to uncover the hidden patterns of concept compatibility within Blake's metaphors by applying a correlation analysis, focusing specifically on the concepts of "Creature" and "Psychological Sphere." The research employs stylometric methods to statistically analyze the frequency and co-occurrence of these concepts within metaphoric structures. The results show that these central concepts are both highly active and productive within Blake's metaphorical framework, with specific tendencies highlighted in how these concepts are combined with others in his poetry.

Silnitsky [22] investigates the English verb system through a differential correlation analysis. The method involves clustering verb features based on their differential correlations, which effectively distinguishes between motivating and limiting functions of verb features. The results demonstrate the complex interrelations within the verb system, where certain features act as dominant factors influencing the behavior of others, thus contributing to the dynamic structure of the English verb system.

Coloma [5] also involves a quantitative linguistic analysis, though its specific details are not fully clear from the provided excerpts. However, it involves the use of correlation analysis to examine the relationships between different linguistic measures. The paper, while less detailed in the provided description, likely follows a similar quantitative approach as in [3] but in a broader linguistic context.

Wahyuni et al. [23] investigate the correlation between students' mastery of parts of speech and their writing achievement among 11th-grade students at MAN 4 AGAM. Using a correlational research design, the authors employed tests to assess parts of speech mastery and documentation to evaluate writing performance. The data analysis, conducted via Pearson Product Moment Correlation using SPSS, revealed a positive low correlation value of 0.376, indicating a statistically significant relationship between the two variables. This suggests that while parts of speech mastery does influence writing achievement, the effect is not substantial.

In [21], the link between mastery of sentence patterns and translation ability among third-semester students at Ahmad Dahlan University is studied. The study employs a quantitative correlational approach with a sample size of 122 students, using sentence pattern and translation tests to measure performance. Pearson’s Product Moment correlation was used for statistical analysis. The results show a moderate positive correlation (r = 0.464) between sentence pattern mastery and translation ability. Students' sentence pattern mastery had an average score of 29.83 (out of 45), with translation abilities averaging 64.50 (out of 100), both classified as moderate.

Köhn et al. [9] explore the relationship between syntax and prosody in spoken German texts. This paper leverages a corpus of 31,803 sentences from the Spoken Wikipedia project, focusing on how syntactic functions, such as subjects and objects, affect prosodic features like pitch, duration, and loudness. The research uses linear mixed effects models to analyze these relationships. Notably, the study found that subject words are spoken with a higher pitch (about 0.2 semitones higher) and longer duration (6% longer), while object words are associated with lower loudness and shorter duration.

Kettunen [8] investigates whether the Type-Token Ratio (TTR) can indicate the morphological complexity of languages. The study analyzes 21 languages using TTR and a Morphological Complexity Index (MCI). Results show that TTR does not consistently correlate with MCI, as languages with high morphological complexity, such as Finnish and Turkish, do not always have high TTR values. Instead, TTR is influenced more by text length and lexical diversity rather than morphological complexity. The study concludes that TTR alone is insufficient for measuring linguistic complexity.

Litvinova [11] discusses the use of computational linguistics for diagnosing personality traits from text through linguistic analysis. The paper highlights the development of a specialized text corpus called "Personality," which includes metatags about authors (gender, psychological traits). The study identifies correlations between grammatical features of texts and authors' characteristics, such as gender and psychological profiles, using natural language processing and statistical modeling. This research has significant applications in forensic linguistics and psychological profiling.

This study builds on foundational research such as [27], which emphasizes that teacher effectiveness is not solely dependent on linguistic features but also involves cognitive and motivational factors. While linguistic features are important, they must be contextualized within broader pedagogical practices. For example, Hattie's meta-analysis identified several non-linguistic traits, such as clarity of instruction and the teacher’s ability to manage classroom dynamics, that contribute to effective teaching.

Furthermore, Hathairat and Maleerat [25] demonstrate the importance of corpus-based vocabulary analysis in the educational context, showing how lexical variety can enhance students' learning. Similarly, Abdul et al. [26] emphasize the need for advanced corpus-based tools for text processing, particularly in multilingual settings. These tools could be adapted to enhance future analyses in this study by improving the identification of key linguistic markers across languages and contexts.

1. PEDAGOGICAL CORPUS AND EXPERIMENTAL STATISTICAL METHODOLOGY

The data for this study consists of automatic transcripts of lessons delivered by secondary school teachers in the Russian Federation. These transcripts were obtained using the Whisper program [18] and subsequently were subject to corrections. Each lesson had an average duration of 40-45 minutes, and the recordings took place in the spring of 2023. The subcorpus of lessons from the more effective teachers contains 33 texts, while the subcorpus of the less effective teachers consists of 22 texts. More effective teachers meet specific criteria such as working in non-selective schools with diverse student populations and achieving above-average results in State Final Certification (the 9th Grade). For the current study foreign language lessons were not included since they are not monolingual and might require a different methodology. The main information about the corpus is presented in Table 1.

Table 1: General Corpus Statistics

| Feature | Value for the More/Less Effective Teachers’ Subcorpus |
| --- | --- |
| Subcorpus size in tokens | 93,841/50,721 |
| Grade 5 texts | 13/8 |
| Grade 6 texts | 7/5 |
| Grade 7 texts | 6/4 |
| Grade 8 texts | 7/4 |
| Grade 9 texts | 0/1 |
| Subjects (for all subcorpora) | Russian, Literature, Biology, History, Geography, Social Studies |

One potential limitation noted in the comparison between more effective and less effective teachers involves the representation of different grades and subjects. While the study attempts to maintain diversity across grades 5–9, future work should consider controlling for subject-specific language patterns. The research [28] suggests that the linguistic demands of different subjects can vary significantly. For instance, history lessons may naturally include more abstract language compared to biology, which often involves concrete terminologies and procedural language. Expanding the corpus to include more lessons from varied subjects will help address this issue. The current study serves as an exploratory investigation rather than a conclusive one.

Linguistic features of the lessons were extracted and quantified using the online Linguistic Profiling Tool [4]. It provides a range of linguistic metrics, such as lexical density, syntactic complexity, and sentence structure, which might be key for identifying distinguishing features between the lessons of more effective and less effective teachers. For each text (lesson), the following linguistic measures were calculated: number of nouns, number of verbs, average length of dependent links, pre- and post-subordinate clauses, sentence length, arity (argument structure complexity), average length of prepositional chains, lexical density, type-token ratio, number of determiners. Once the linguistic values were collected, the distribution of each value samples were analyzed to determine whether they followed a normal or abnormal distribution (Smirnov-Kolmogorov test of normality). Next, intertextual correlations between different pairs of linguistic features were calculated. The aim was to explore how syntactic, lexical, and structural elements relate to one another across the different lessons. The combination of such correlations are *linguistic pedagogical profiles*, i.e. a set of linguistic correlates that characterize lesson features. This is a modified definition of *linguistic profiles* described in [12, 13]. While factors such as social status and student-teacher relationships undoubtedly influence educational outcomes, they fall outside the scope of this particular research, which is designed to quantify and analyze linguistic patterns alone. Additionally, the absence of data on social dynamics makes it impossible to factor them into this analysis. By isolating linguistic features, this study provides an essential first step toward understanding how language shapes pedagogical effectiveness, which can later be supplemented with broader contextual factors in future research. The following correlations were computed:

1. Nouns-Verbs.
2. Dependent Links-Average Sentence Length.
3. Pre-Subordinate Clause-Average Sentence Length.
4. Post-Subordinate Clause-Average Sentence Length.
5. Average Arity-Average Sentence Length.
6. Average Prepositional Chain Length-Average Sentence Length.
7. Lexical Density-Type Token Ratio (TTR).
8. Number of Determiners-Average Sentence Length.

The choice of these correlations is based on theoretical background provided in [1, 10, 15, 17, 20]. For instance, if there are few nouns used in the *nouns-verbs* correlation, it indicates a tendency towards using a nominative structure in the construction of pedagogical discourse or the predominance of predicate structures. The value of the *lexical density-type token ration* correlation investigates the relationship between lexical density (vocabulary richness) and the type-token ratio (variation in vocabulary usage). Finally, as for the *average arity-average sentence length* correlation, verbs with higher arity typically require more explicit arguments or complements. A transitive verb (arity of 2) may lead to longer sentences because it necessitates a subject and a direct object. In contrast, intransitive verbs (arity of 1) usually require fewer components.

These correlations were calculated using appropriate mathematical formulae to obtain intertextual correlation values for each linguistic feature pair. Given that the current datasets tend to deviate from a normal distribution, Spearman's rank-order correlation was chosen for the analysis. The Spearman correlation was calculated using the Correlations Widget in the Orange Data Mining tool [6]. P-values were computed to assess whether the observed correlations were statistically significant, i.e., whether they could be generalized beyond the sample data. Significance levels were set at α = 0.05. Finally, to visualize the relationships between linguistic features, scatter plots were generated using the Scatter Plot Widget in Orange Data Mining. These scatter plots illustrated the strength and direction of the correlations.

1. RESULTS AND DISCUSSION

Table 2 shows the resultant linguistic profiles for both subcorpora presented in the form of correlation values.

Table 2: Pedagogical Linguistic Profiles

| Linguistic Correlations | More Effective Teachers | Less Effective Teachers |
| --- | --- | --- |
| Nouns-Verbs | -0.202 | -0.255 |
| Dependent Links-Average Sentence Length | 0.908 | 0.88 |
| Pre-Subordinate Clause-Average Sentence Length | 0.076 | -0.011 |
| Post-Subordinate Clause-Average Sentence Length | -0.076 | 0.011 |
| Average Arity-Average Sentence Length | 0.789 | 0.78 |
| Average Prepositional Chain Length-Average Sentence Length | 0.131 | -0.401 |
| Lexical Density-Type Token Ratio | 0.185 | 0.267 |
| Number of Determiners-Average Sentence Length | 0.341 | 0.3 |

The linguistic correlations shown in Table 2 represent key metrics used to assess differences between more and less effective teachers. The values provide a comparative overview of linguistic tendencies, but they are not intended to serve as formal markers of effectiveness across all subjects and grades. Instead, they offer insight into broad trends in teaching language use, highlighting general linguistic behaviors that may influence effectiveness, regardless of specific content or grade level. It should be mentioned that the *dependent links-average sentence length, average arity-average sentence length, number of determiners-average sentence* length correlations cannot be considered formal markers of determining whether a particular teacher is more or less effective since the values are almost the same.

* 1. Nouns-Verbs

The two visualizations (Figure 1) display correlations between the use of nouns and verbs for two groups of teachers: more effective (ВР) and less effective (НР). Both groups show negative correlations between noun and verb usage, with the ВР group having a slightly less negative correlation (r = -0.202) compared to the НР group (r = -0.255). In the more effective group, the data points are spread more evenly across different noun and verb usage combinations. There is a noticeable mix of both high and low noun/verb densities, indicating a diverse approach to language, potentially pointing to a flexible and adaptive teaching style. In contrast, the less effective group displays a clearer pattern of clustering, with a more distinct negative relationship between noun and verb use. This could indicate an approach where teachers tend to overuse either nouns or verbs but not both simultaneously. This imbalance may lead to less engaging or clear instruction. The comparison of teaching practices and their linguistic interpretation is provided in Table 3.

Table 3: Examples for the Nouns-Verbs correlation

| Russian Example | English Translation | Linguistic Interpretation |
| --- | --- | --- |
| More Effective Teachers | | |
| Pravil'no, ***zapisyvaj***, uzhe ***zapisyvat'*** nado. Uzhe vse ***nachertili***, vrode kak. (History, the 5th Grade) | That's right, ***write*** it ***down***, it's time to ***write*** it ***down***. We ***have*** already ***drawn*** everything, it seems (History, the 5th Grade). | No nouns are used. Teaching recommendation |
| Immunitet. ***Davajte***, ***davajte***, ***davajte***, shire ***dumajte***, ***dumajte***, ***dumajte***, ***davajte***, ***ne stesnyajtes'***, vot kakie u vas associacii prihodyat – ***govorite***. (Biology, the 8th Grade) | Immunity. ***Come on***, ***come on***, ***come on***, ***think*** outside the box, ***think***, ***think***, ***come on***, ***don't be shy***, what associations ***come*** to you – ***tell*** me (Biology, the 8th Grade). | Very few nouns are used. Lexical repetition of verbs denotes a very active call for action |
| Less Effective Teachers | | |
| ***Kupechestvo*** postepenno prevrashchaetsya v ***burzhuaziyu***. Zapisyvaem. ***Kupechestvo*** postepenno prevrashchaetsya v ***burzhuaziyu***. Vsyo, to est' vot u nas, pozhalujsta, s vami rabochij ***klass***, ***burzhuaziya***, ***predpriyatiya***. (Social Studies, the 8th Grade) | The ***merchant class*** is gradually turning into ***the bourgeoisie***. Let's write it down. The ***merchant class*** is gradually turning into ***the bourgeoisie***. That's it, that is, here we have, please, ***the*** working ***class***, ***the bourgeoisie***, ***the enterprises*** (Social Studies, the 8th Grade). | Parallel syntactic constructions. Repetition of noun terms without a description of what contribution they make to a given fragment of the lesson indicates uncertainty in the pedagogical speech. |
| Vpolne veroyatno. To est', smotrite. Ya vam chasto govoryu, chto, devochki, chto ya vam govoryu? Kogda vy ***vra… vyrastite***, vy ***b… obrashchajte*** vnimanie na… na chto? (Literature, the 6th Grade) | Probably. I mean, look. I often tell you that, girls, what do I tell you? When you ***gr… grow*** up, you ***p... pay*** attention to... to what? (Literature, the 6th Grade) | Hesitation on verbs indicate the vague formulation of pedagogical thoughts. |

Table 3 provides concrete examples of noun-verb usage from both more and less effective teachers across different subjects and grades. These examples illustrate typical linguistic patterns but are not meant to imply a strict correspondence between specific subjects or grades and teacher effectiveness. Rather, the table demonstrates how less effective teachers tend to show rigid or unbalanced language patterns. This distinction holds across subjects, focusing on language use rather than the content of specific disciplines.

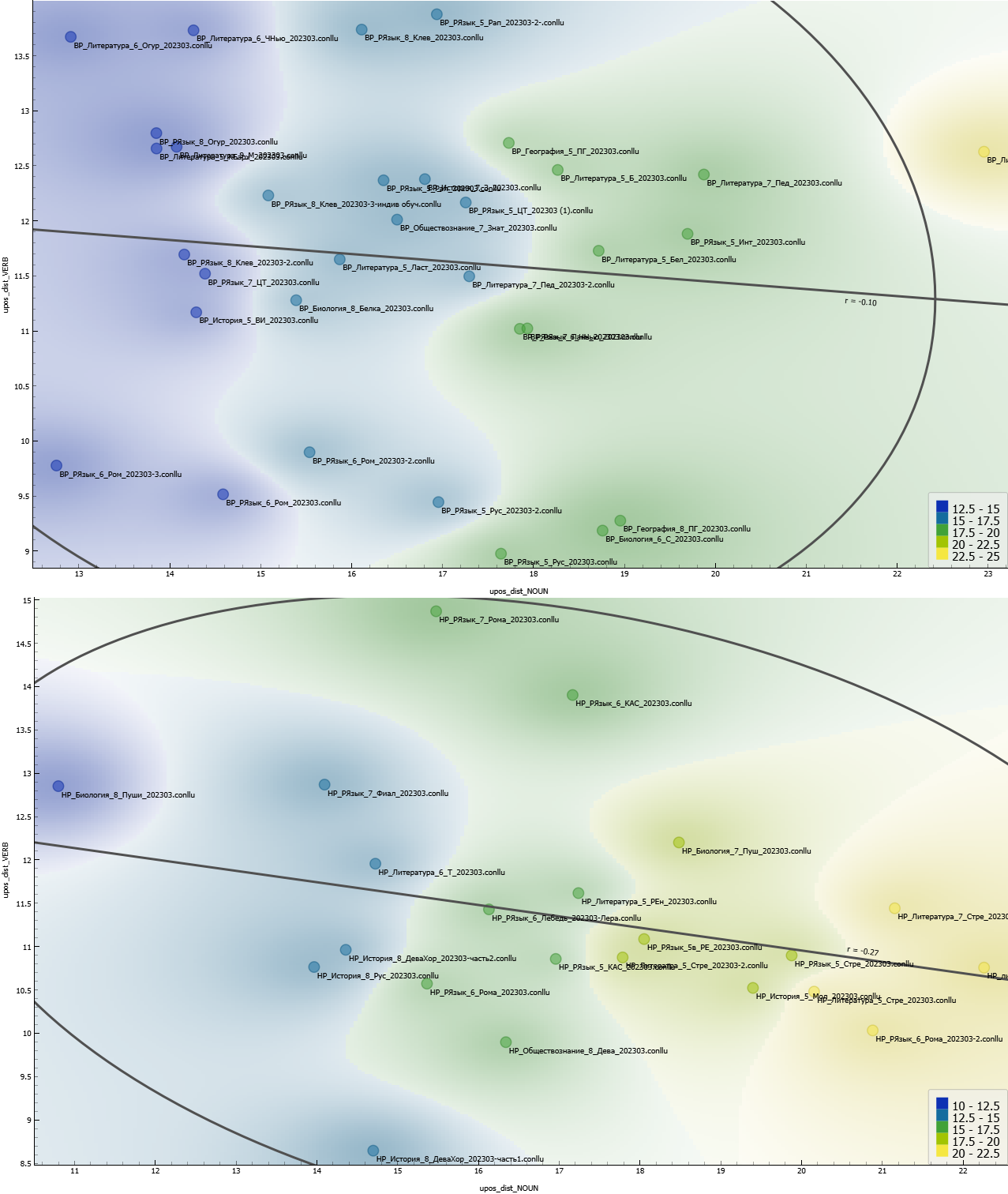


Figure 1: The visualization of the *nouns-verbs* correlation for more effective (above) and less effective (below) teachers

* 1. Lexical Density-Type Token Ratio

The correlation between lexical density and TTR is weaker (Figure 2). It is suggested that more effective teachers use a wider variety of unique words without significantly increasing lexical density. The points for subjects like literature and biology show a mix of lexical strategies, with content density ranging from high to low, which may suggest these teachers adapt their speech depending on the subject matter.

The correlation is slightly stronger in the less effective group, implying that an increase in lexical density is more closely tied to an increase in unique lemmata. The distribution is more compact, and the points tend to cluster at moderate-to-low lexical density values, with TTR moderately varying. Less effective teachers might prioritize vocabulary diversity (for example, using a plethora of terms), but this rigidity could result in more complex language that might not always suit the comprehension needs of their students. Some examples are presented in Table 4.

Table 4: Examples for the Lexical Density-Type Token Ratio correlation

| Russian Example | English Translation | Linguistic Interpretation |
| --- | --- | --- |
| More Effective Teachers | | |
| Potomu chto korni, rastushchie ot ***steblej*** ili ot ***list'ev***, nazyvayutsya ***pridatochnymi***. Esli ya ego posazhu v otdel'nyj gorshochek, u menya vyrastet ***rasteniya***? (Biology, the 6th Grade) | Because ***roots*** that grow from ***stems*** or ***leaves*** are called ***adventitious***. If I plant it in a separate pot, will I grow a ***plant***? (Biology, the 6th Grade) | The usage of five terms per 24 words. There are no repetitions of terms. |
| Less Effective Teachers | | |
| ***Kryl'ya*** – eto ***perednie konechnosti***, kotorye prevratilis' v ***kryl'ya***, da, to est' esli my ***zemnovodnyh*** budem sravnivat' s ***pticami***, tam dve ***lapki*** dve pary ***lapok*** vernee. (Biology, the 7th Grade) | ***Wings*** are the ***forelimbs*** that have turned into ***wings***, yes, that is, if we compare ***amphibians*** with ***birds***, there are two ***feet***, it is better to say that they have two pairs of ***feet***. (Biology, the 7th Grade) | The usage of seven terms per 25 words. Repeating terms inside definitions might confuse students. |

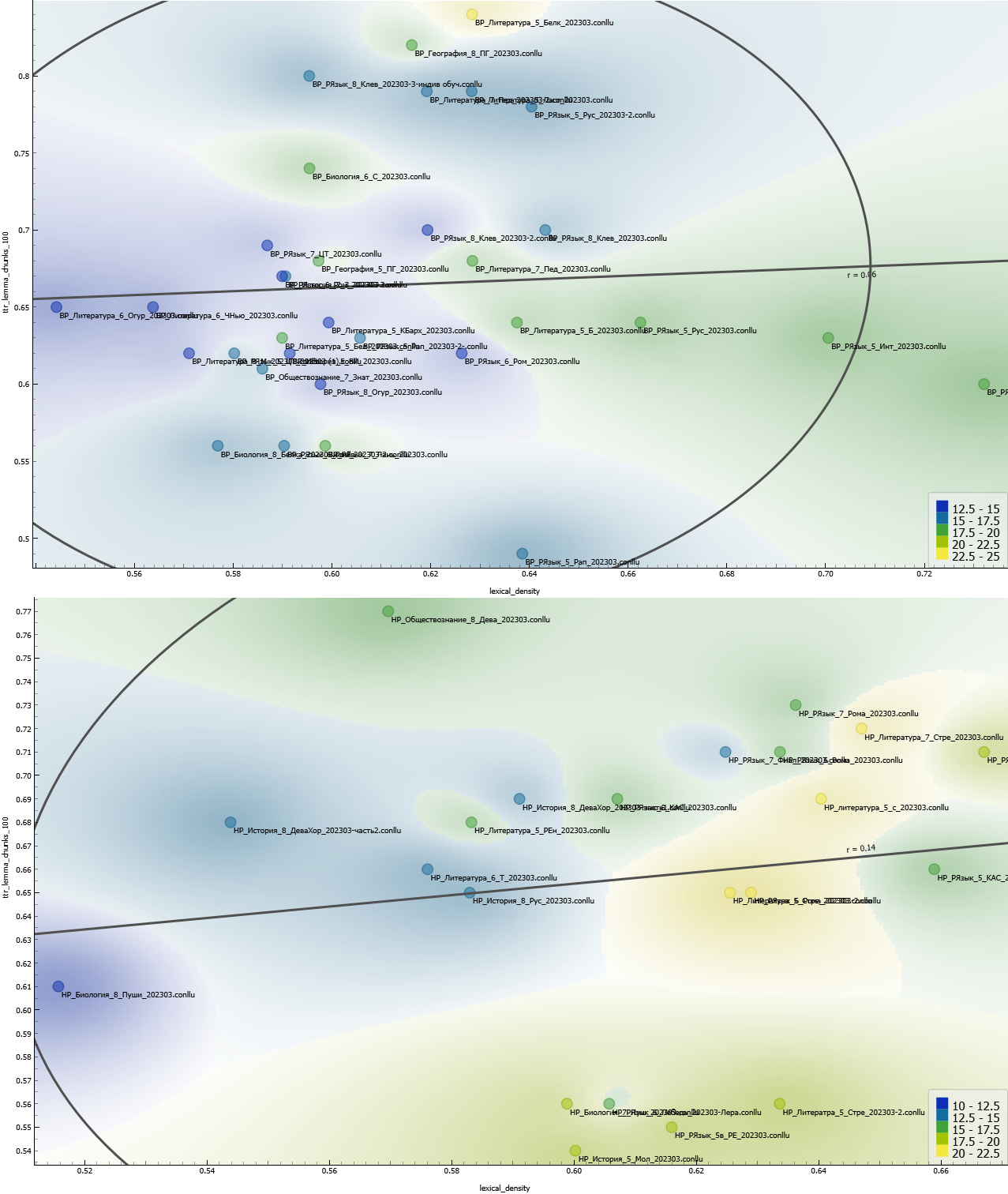


Figure 2: The visualization of the *lexical density-type token ratio* correlation for more effective (above) and less effective (below) teachers

* 1. Average Prepositional Chain Length-Average Sentence Length

**Figure 3 presents two scatter plots for correlation. More effective teachers** show a slight positive correlation (r = 0.131). In the lower left corner of the upper plot, we observe a concentration of points with both short prepositional chains and shorter sentences. However, moving toward the upper right, we see points with longer chains and longer sentences, which reinforces the trend. The overall situation reflects better command of complex sentence construction and suggests more fluid, sophisticated speech patterns in the classroom. On the other side, there is a stronger negative correlation (r = -0.401) for less effective teachers. A more notable cluster of points in the upper left part of the bottom plot suggests a frequent usage of longer sentences with shorter prepositional chains. As prepositional complexity increases (rightward movement), we see fewer points with long sentences, confirming the negative correlation. It is suggested that teachers’ speech becomes more fragmented or less cohesive when using complex syntactic structures, which might impact the clarity of their communication. Some examples are presented in Table 5.

Table 5: Examples for the Average Prepositional Chain Length-Average Sentence Length correlation

| Russian Example | English Translation | Linguistic Interpretation |
| --- | --- | --- |
| More Effective Teachers | | |
| Eto proizvedenie bylo napechatano ***v samom nachale*** tvorcheskoj deyatel'nosti Maksima Gor'kogo, kogda on eshchyo ne byl znamenitym pisatelem. No poyavilos' ono ***v pechati***, vyshlo ***v svet v «Samarskoj gazete» v tysyacha vosem'sot devyanosto pyatom godu***, kogda on byl uzhe yarkim i izvestnym pisatelem. (Literature, the 7th Grade) | This work was published ***at the very beginning*** of Maxim Gorky's creative activity, when he was not yet a famous writer. But it appeared ***in print***, ***came out (a prepositional construction in the Russian language)*** ***in "Samara newspaper" in one thousand eight hundred and ninety-five***, when he was already a bright and famous writer. (Literature, the 7th Grade) | The usage of extended prepositional constructions in compound and complex sentences. |
| Less Effective Teachers | | |
| Sosloviya, molodcy, pravil'no, sosloviya, da? To est', eto vsegda ***u nas*** bylo. Takzhe ***u nas***, poluchaetsya, chto byli? Potomstvennye i lichnye dvoryane, da? Pochyotnye grazhdane, kupcy. U kupcov po… vspominaem, vspominaem, ch… kupcy kak delyatsya? **Na kakie kategorii**? ***U kupcov*** kakie kategorii? Shest'desyat vos'maya stranichka, podsmatrivaem, podskazyvaem sebe. (Social Studies, the 8th Grade) | Estates, well done, right, estates, right? That is, ***we (a prepositional construction in the Russian language)*** have always had this. What did ***we (a prepositional construction in the Russian language)*** turn out to have? Hereditary and personal nobles, right? Honorary citizens, merchants. Merchants have… we remember, we remember, how are merchants divided? ***Into what categories***? What categories do merchants have? Page sixty-eight, look, prompt yourselves. (Social Studies, the 8th Grade) | The usage of prepositional constructions in short simple sentences, most of the constructions have the head (a preposition) and a dependent (a noun), they are not extended. |

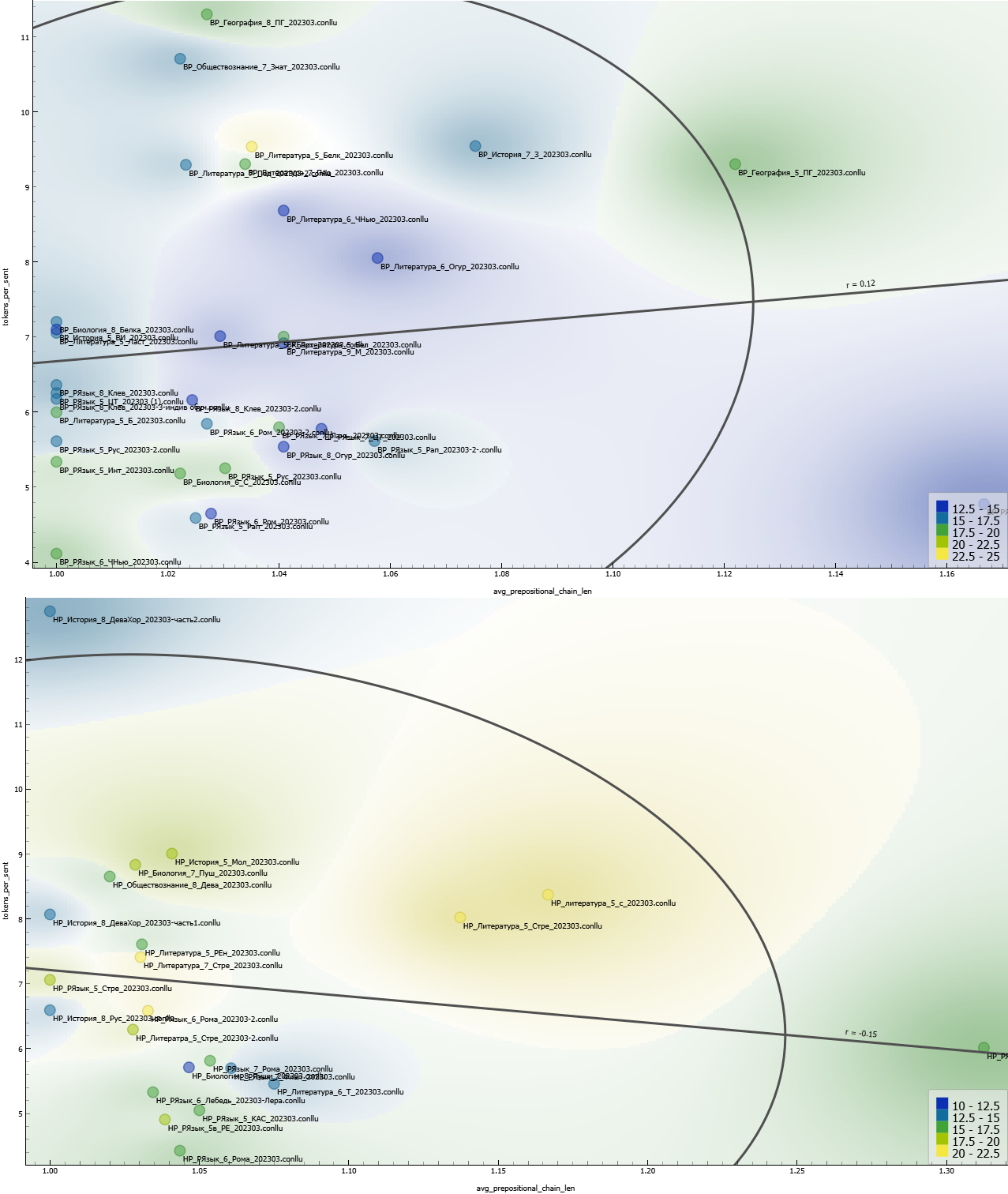


Figure 3: The visualization of the *average prepositional chain length-average sentence length* correlation for more effective (above) and less effective (below) teachers

* 1. Pre-Subordinate Clause-Average Sentence Length

Figure 4 shows that the data points for more effective teachers are more dispersed across both the subordinate clause axis and the average sentence length axis, indicating a broader variability in both clause usage and sentence complexity among these teachers. As for less effective teachers, there is a more concentrated clustering of data points around shorter sentence lengths and lower pre-subordinate clause usage. Despite all the values being almost equal to zero, less effective teachers tend to use simple and straightforward sentences. If we turn to the *post-subordinate clause-average sentence length* correlation, the results are the same, but with the opposite sign. Table 6 presents some examples (proper nouns of students are anonymized with X).

Table 6: Examples for the Pre-Subordinate Clause-Average Sentence Length correlation

| Russian Example | English Translation | Linguistic Interpretation |
| --- | --- | --- |
| More Effective Teachers | | |
| ***Esli by vas poprosili v kraskah izobrazit' etu kartinu***, kakie by cveta vy ispol'zovali, X? (Literature, the 5th Grade) | ***If you were asked to paint this picture***, what colors would you use, X? (Literature, the 5th Grade) | Using a conditional sentence with a pre-subordinate clause makes students pay attention to the topic of the lesson. |
| Da, ***kto vse sdelal, poka zhdyom ot drugih***, mozhete i chislo napisat'. (Russian, the 5th Grade) | Yes, those, ***who did everything***, can write the date ***while we are waiting for the others (both clauses are pre-subordinate in the Russian language)***. (Russian, the 5th Grade) | The use of subordinate clauses allows the teacher to parallelize the actions of students in the classroom. |
| Less Effective Teachers | | |
| Znachit, i ona v tysyacha sem'sot sorok chetvyortom godu privozit emu iz Germanii, mozhno skazat', vypisyvaet, vot… zvuchit eto dostatochno grubo, da? (History, the 8th Grade) | So, in 1744, she brings it to him from Germany, you could say, orders it, well... it sounds pretty crude, right? (History, the 8th Grade) | No pre-subordinate clauses are used. Paraphrasing the sentence makes it difficult to follow the main idea of the lesson. |
| Vot sejchas my pokazyvaem. Tak. Horosho. My s vami eshchyo posmotrim migraciyu ptic segodnya, nebol'shoj interaktiv. Na primere polyarnoj pticy, kotorye u nas obitayut v Arktike. (Biology, the 7th Grade) | We are showing it now. So. Great. We will also look at bird migration today, it will be a small interactive part of the lesson. Using the example of polar birds that live in the Arctic. (Biology, the 7th Grade) | No pre-subordinate clauses are used. Simplifying compound sentences through the use of parcellation. |

More effective teachers tend to use more complex sentence structures, which can promote deeper student engagement, while less effective teachers may simplify their language. Teachers who are less experienced or lack confidence in their teaching abilities may default to more straightforward language to avoid making mistakes or to ensure clarity. Studies like [16] show that teachers modify their speech when they feel uncertain about their students' proficiency, often simplifying their language to reduce complexity in the classroom.

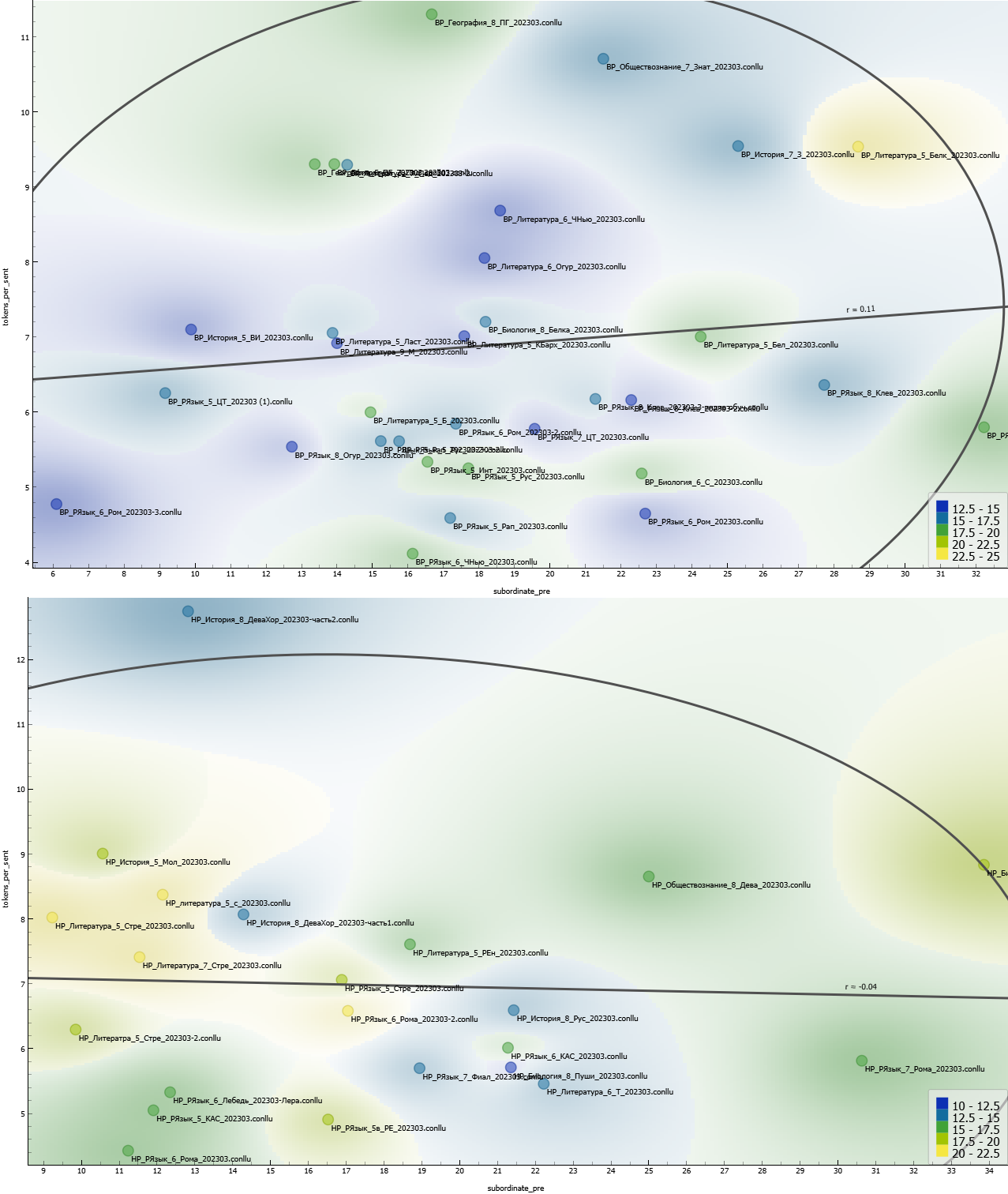


Figure 4: The visualization of the *pre-subordinate clause-average sentence length* correlation for more effective (above) and less effective (below) teachers

The analysis reveals that more effective teachers tend to use a wider variety of syntactic structures and maintain a balanced noun-verb ratio, indicating flexibility in their instructional language. This variety likely enhances student engagement by making the lesson content more dynamic and accessible. In contrast, less effective teachers exhibit more rigid, repetitive patterns with a tendency to overuse either nouns or verbs, which can result in a less engaging and more monotonous discourse. Moreover, effective teachers display a more sophisticated use of prepositional chains and sentence complexity, suggesting they can construct more nuanced explanations. Their ability to adjust language complexity, perhaps by adapting to students’ cognitive needs, aligns with research showing that pedagogical success involves language [27]. These linguistic traits, therefore, not only reflect teaching style but also likely contribute to more effective classroom communication and learning outcomes.

1. Conclusion and future work

This research continues a series of studies on effective and ineffective pedagogical practices [2, 7, 16]. In this study, we have successfully developed linguistic pedagogical profiles for more effective and less effective teachers using statistical and machine learning methods. The analysis demonstrated that key linguistic features – such as noun-verb usage, sentence complexity, and lexical density – correlate with teaching effectiveness. More effective teachers were found to employ a balanced and flexible linguistic approach, characterized by varied syntactic structures and a broader range of lexical items. In contrast, less effective teachers exhibited more rigid and repetitive patterns, which may hinder student engagement and comprehension. These findings provide an important step towards understanding how linguistic factors contribute to successful pedagogy.

However, the current study is not without limitations. The corpus size, while sufficient for initial analysis, could be expanded to include a more diverse range of subjects and educational contexts. Additionally, while the correlations identified in this research highlight important trends, they do not fully capture the dynamic interaction between linguistic features and teaching outcomes. Future studies could address these limitations by employing larger datasets and incorporating qualitative methods, such as student feedback, to better understand the nuances of effective teaching.

Looking forward, there are several promising directions for further research. One of them is the application of these linguistic profiles in teacher training programs, where they could serve as diagnostic tools for improving pedagogical practices. Additionally, expanding this research to include cross-linguistic and cross-cultural comparisons would provide valuable insights into the universality of the identified linguistic features in different educational systems. Finally, the integration of real-time linguistic analysis through AI-powered tools could offer immediate feedback to teachers, helping them adjust their language use to enhance lesson effectiveness.

ACKNOWLEDGMENTS

The author acknowledges Saint Petersburg State University for a research project 103923108 and SberObrazovanie LLC for a research No. 230712-107-YUL.

REFERENCES

1. Akimova G. N., Vyatkina S. V., Kazakov V. P., Rudnev D. V. 2013. Syntax of the modern Russian language. Faculty of Philology, St. Petersburg State University.
2. Alibali, M. V., Nathan, M. J., Wolfgram M. S., Breckinridge R. C., Jacobs S. A., Martinez C. J. and Knuth E. J. 2014. How Teachers Link Ideas in Mathematics Instruction Using Speech and Gesture: A Corpus Analysis. Cognition and Instruction. Vol. 32, № 1, 65-100.
3. Baburchenkova, I. O. 2016. Compatibility of Units for the Conceptual Sphere in the Early Works of W. Blake: Correlation Analysis [Sochetaemost' edinic konceptosfery v rannem tvorchestve U. Blejka: korrelyacionnyj analiz]. Quantitative Linguistics, Vol. 4, 31-43.
4. Brunato, D., Cimino, A., Dell’Orletta F., Venturi, G., and Montemagni, S. 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts. Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association, 7145–7151.
5. Coloma, G. 2022. Correlation between Linguistic Measures: An Extended Analysis. Studies in Linguistics and Literature, Vol. 6, 109–132. <http://dx.doi.org/10.22158/sll.v6n4p109>.
6. Demšar, J., Zupan, B., Leban, G., and Curk, T. 2004. Orange: From Experimental Machine Learning to Interactive Data Mining. Lecture Notes in Computer Science, 537–539. <http://dx.doi.org/10.1007/978-3-540-30116-5_58>.
7. Ford, C. E. 1984. The Influence of Speech Variety on Teachers' Evaluation of Students with Comparable Academic Ability. TESOL Quarterly. Vol. 18, № 1, 25-40.
8. Kettunen, K. 2014. Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? Journal of Quantitative Linguistics 21, 223–245. <http://dx.doi.org/10.1080/09296174.2014.911506>.
9. Köhn, A., Baumann, T., and Dörfler, O. 2018. An Empirical Analysis of the Correlation of Syntax and Prosody. <https://arxiv.org/abs/1806.05900>.
10. Konyushkevich M. I. 2013. Transformation of Prepositional-Case Syntaxeme into a Predicative Unit: Correlation of Preposition and Connection Indicator of a Complex Sentence. Linguistic Studies, Vol. 26, 93–99.
11. Litvinova, T. A. 2015. Resources of Computational Linguistics for the Task of Author Profiling: a Corpus-Based Study. Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication, Vol. 3, 37-41.
12. Mamaev, I. D. 2024. Cluster Analysis of Linguistic Profiles of Hidden Communities. Philology. Theory and Practice, Vol. 17, №5, 1739–1747. <http://dx.doi.org/10.30853/phil20240250>.
13. Mamaev, I. D. 2024. Linguistic Profiles of Hidden Communities: A Morphosyntactic Aspect. Philology. Theory and Practice, Vol. 17, №4, 1155–1162. <http://dx.doi.org/10.30853/phil20240168>.
14. Martin-Kniep G. O. 2007. Communities That Learn, Lead, and Last: Building and Sustaining Educational Expertise. John Wiley & Sons.
15. Orekhov, B.V. 2022. Metric and lexical Diversity in the Poems of A.A. Voznesensky. Proceedings of the V.V. Vinogradov Russian Language Institute, 50–58. <http://dx.doi.org/10.31912/pvrli-2022.3.4>.
16. Owen, D. 1996. Do Teachers Modify Their Speech According to the Proficiency of their Students. English Language Teacher Education and Development, Vol. 2, №. 1, 31-51.
17. Potebnya A. A. 1958. Notes on Russian Grammar [Iz zapisok po russkoj grammatike]. Uchpedgiz Publisher, Vol. 1–2.
18. Radford, A., Kim J. W., Xu T., Brockman G., McLeavey C., and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. Proceedings of the 40th International Conference on Machine Learning (ICML'23). Vol. 202, Article 1182, 28492–28518.
19. Rasinger, S. M. 2013. Quantitative Research in Linguistics: An Introduction. A&C Black.
20. Reinhart, T. and Siloni, T. 2005. The Lexicon-Syntax Parameter: Reflexivization and Other Arity Operations. Linguistic Inquiry 36, 389–436. <http://dx.doi.org/10.1162/0024389054396881>.
21. Sari, D. M. 2022. The Correlation Between Sentence Patterns Mastery and Translating Ability. Zenodo. Vol. 8, №16, 163–170.
22. Silnitsky, G. G. 2017. Differential Correlation Analysis of Multilevel Verb Features in English. Proceedings of Smolensk State University, Vol. 4, 197–210.
23. Wahyuni, C. and Safitri, L. 2022. The Correlation Between Parts of Speech Mastery and Writing Achievement. International Journal of Research on English Teaching and Applied Linguistics, Vol. 2, 37–44. <http://dx.doi.org/10.30863/ijretal.v2i2.2453>.
24. Walton, J., Priest, N., and Paradies, Y. 2013. Identifying and Developing Effective Approaches to Foster Intercultural Understanding in Schools. Intercultural Education, Vol. 24, 181–194. <http://dx.doi.org/10.1080/14675986.2013.793036>.
25. Ketmaneechairat, H., and Maliyaem, M. 2023. Corpus-Based Vocabulary List for Thai Language. Journal of Advances in Information Technology, Vol. 14, №2, 319-327, 2023.
26. Jabbar, A., Illahi, M., Iqbal, S., Khan, A. R., ElHakim, N., and Saba, T. 2023. PWMStem: A Corpus-Based Suffix Identification and Stripping Algorithm for Multi-lingual Stemming, Journal of Advances in Information Technology, Vol. 14, №4, 863-875. <http://dx.doi.org/10.12720/jait.14.4.863-875>.
27. Hattie, J. 2003. Teachers make a difference: What is the research evidence? *Paper presented at the Australian Council for Educational Research Annual Conference on Building Teacher Quality*. Melbourne, Australia. Retrieved from <http://research.acer.edu.au/research_conference_2003/4/>.
28. Redish, E. F., and Kuo, E. 2015. Language of physics, language of math: Disciplinary culture and dynamic epistemology. Science & Education, Vol. 24, 561-590.