Communication Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 303–311
DOI: 10.15439/2023F7698
ISSN 2300-5963 ACSIS, Vol. 37

The Grammar and Syntax Based Corpus Analysis Tool for the Ukrainian Language

Daria Stetsenko
0000-0002-3698-4414
NASK National Research Institute
Kolska 12, 01-045 Warsaw, Poland
Email: {daria.stetsenko@nask.pl}

Inez Okulska
0000-0002-1452-9840
NASK National Research Institute
Kolska 12, 01-045 Warsaw, Poland
Email:{inez.okulska@nask.pl}

***Abstract*—This paper provides an overview of a corpus analysis tool - the StyloMetrix for the Ukrainian language. The StyloMetrix incorporates 104 metrics that cover grammatical, stylistic, and syntactic patterns.**

The idea of constructing the statistical evaluation of syntactic and grammar features is straightforward and familiar for the languages like English, Spanish, German, and others; it is yet to be developed for low-resource languages like Ukrainian. We describe the StyloMetrix pipeline and provide some experiments with this tool for the text classification task. We also describe our package's main limitations and the metrics' evaluation procedure.

***Index Terms*—stylometric analysis, Ukrainian linguistics metrics, text analysis, supervised learning.**

I. Introduction

Ukrainian remains one of the low-resource languages with few practical applications in machine learning and deep learning. Many studies on the Ukrainian language are conducted in terms of multilingual settings, such as training the multilingual large language models [14], [18], transformers [23], [6], or abstractive summarization [10]. We offer a corpus analysis tool for the Ukrainian language – the StyloMetrix. The underlying idea is not new in the NLP community but is recent in the Ukrainian language.

This paper provides an overview of an open-source Python package – the StyloMetrix developed initially for the Polish language and further extended for English and recently for Ukrainian. The StyloMetrix is built upon a range of metrics crafted manually by computational linguists and researchers from literary studies to analyze stylometric features of texts from different genres. The principal purport of this package is to provide high-quality statistical evaluations of the general grammatical, lexical, and syntactic features of the text, regardless of its length, genre, or author.

We organize our paper in the following way:

- we provide an overview of similar tools for text analysis and a general idea of the corpus linguistics based on the syntactic and grammar representations;
- give an exhaustive characteristic of existing metrics for the Ukrainian language, their evaluation, and limitations;
- describe a case study with the StyloMetrix as the baseline model for the text classification task, providing the metrics analysis and feature importance of the classification model.

II. Related Studies

The idea to measure specific textual features to determine a text's register or an author is not new. In 1998, D. Biber, S. Conrad and R. Reppen have developed a comprehensive methodological approach for corpus analysis based only on grammatical characteristics. D. Biber argues that, although, semantic evaluations and descriptive analysis can provide a valuable insight about the narrative, it is not enough if one needs to discern the genre of the text or to assess whether it belongs to a particular author and an epoch [4]. On the other hand, grammatical/syntactic characteristics and figures of speech may come in handy and be less decisive and more exhaustive when it comes to genre, author or style estimation. M.A.K. Halliday supports this view and emphasizes the general importance of corpus studies as a source of insight into the nature of language. He points out that *a language is inherently probabilistic and we need to extract the frequencies in the texts to establish probabilities in the grammatical system – not for the purpose of tagging and parsing, but to discover the interactions between different subsystems* [2].

The development of corpus-based grammar and syntactic tools for text mining has started in 1990s and is still an ongoing field of investigation. Some of the corpus-based techniques aim to manually study the English grammar and discourse. For instance, [1] and [16] provide introductions on how to identify and extract syntactic and grammatical constructions in corpora to build tagging and parsing algorithms. They cover various aspects, limitations and boundaries related to grammar and syntax. Other researchers concentrate on specific incarnations of the language use. For example, [29] on negation and lexical diffusion in syntactic change; [8] on prosody and pragmatics based on it-clefts and wh-clefts; [12] on automated retrieval of passives; [16] on infinitival complement clauses; and [7] has conducted the most valuable

https://github.com/ZILiAT-NASK/StyloMetrix

Thematic track: International Workshop on AI in Digital Humanities, Computational Social Sciences and Economics Research

study on generative grammar that has served as a scaffold for contemporary natural language processing. Those techniques are the basis of modern tools and web-based services for text analysis.

We follow the assumption that grammar and syntax can be enough for the tasks connected with style and author classification which are unified under the term stylometry [20].

The most popular applications for the stylometric analysis are the "Stylometry with R" (stylo) [9] (for English and Polish), WebSty [15] and CohMetrix [11]. The stylo is a flexible R package for the high-level analysis of writing style in stylometry. The package can be applied at the supervised learning for the text classification [9]. WebSty [15] is an accessible open-sourced library that encompasses grammatical, lexical, and thematic parameters which can be manually selected by the user. The tool covers the Polish, English, German and Hungarian languages. Coh-Metrix is a web-based platform that offers a wider range of descriptive statistic measurements. For example, low-level metrics counting pronouns per sentence, Text Easability Principal Component Scores, Referential Cohesion, LSA, Lexical Diversity, Connectives, Situation Model, Syntactic Complexity, Syntactic Pattern Density, Word Information, Readability, etc. [17]. The documented versions of Coh-Metrix exist for Spanish [24], Portuguese [25], and Chinese [22] (however, they are developed independently and not supported by the initial authors).

There are many tools for corpus analysis that look at concordances, n-grams, co-locations, key words and numerous frequency analysis which can be applied for the stylometric classification tasks, but most of them are quite primitive and basic with respect to the intricacy of grammar structures like tenses or syntactic phrases (the comprehensive list of tools can be accessed via the link `https://corpus-analysis.com/`

Therefore, inspired by the powerful image of grammatical patterns and syntactic clauses we build the first (to our knowledge) corpus-analysis tool for the Ukrainian language that presents a thorough statistical evaluation of the Ukrainian grammar, syntactic patterns, and some descriptive lexical assessment.

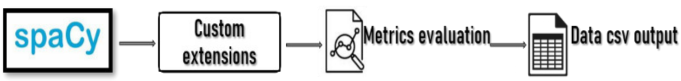


Fig. I The pipeline of the StyloMetrix.

III. Grammatical Vector Representations

A. General outline

The general pipeline of the tool is presented in the Fig. I. First, we utilize the standard spaCy pipeline of the transformer model for the Ukrainian language. The primary purpose of our package is not to build new tagger or parser algorithms but to add a higher level of grammatical and syntactic language characteristics and provide descriptive statistical measurements for each of them. Ukrainian is a fusional language, and the basic spaCy pipeline `https://spacy.io/models/uk` can trace only primary morphological features such as animacy, gender, case, number, aspect, degree, name type, verb form, and others. Nonetheless, these attributes do not cover all aspects of Ukrainian morphology and grammar, such as two types of conjugation, four types of declension, and present, past, and future tenses. Therefore, we leverage the last spaCy component of the pipeline and create custom extensions for each case. Further, the tokens that fall under specific rules are calculated by the discussed formula at the stage of the Metric evaluation and are stored in the data frame that is further available for a user in the .csv format. As for the input – the StyloMetrix can be applied to any text length starting from a single sentence.

The StyloMetrix is a tool designed to calculate the mean value of a distinct grammar rule, a lexical component or a stylistic phenomenon. The statistical measurement is derived by the standard formula:

$$\frac{\sum_0^n w}{\mathrm{N}}$$

Where $\sum_0^n w$ is the sum of all tokens that fall under the particular rule, and N – is the total amount of tokens in the text. This evaluation holds for all metrics. Hence the output is acquired as a matrix, where the text instances are at the y-axis and the x-axis is the vectors of real numbers that stand for the specific metric. The obtained matrices can be utilized for various machine-learning tasks.

Primary developed for the Polish language, which is also fusional, the package has been used for stylometric analysis and text classification. For example, [21] presents a study on erotic vs. neutral text classification using the StyloMetrix vectors as the input to the RandomForest Classifier. The general accuracy has yielded around 0.90 score, giving us an impetus to deliver the primary metrics for the Ukrainian language and test their performance on the existing annotated datasets.

B. Spacy Limitations

Before developing the rules for custom extensions and metrics, we verified the spaCy tags' correctness. Table I presents the incongruencies which have been discerned. Among the most frequent mis assignments are morphological features such as case, animacy, aspect, and gender. For example, "Ілля" is a typical male Ukrainian name that is

TABLE I.
SPACY TAGS INCONGRUENCIES.

word	spacy	Correct tag	Sentence
закрапало	Aspect=Imp	Aspect=Perf	Із стріх закрапало, а з гір струмочки покотилися.
веснянки	ADV	NOUN, Plural	Вже веснянки заспівали.
замазалося	Aspect=Imp	Aspect=Perf	Високе небо замазалося зеленобурими хмарами, припало до землі, наче нагнітило на неї.
крук	Animacy=Inan	Animacy=Animate	Тільки чорний крук надувся, жалібно закрякав з високої могили серед пустельного поля.
завдання	Case=Acc	Case-Nom	Завдання буде зроблено.
листа	Animacy=Anim	Animacy=Inan	Я напушу листа.
осінню	ADJ Case=Acc	NOUN Case=Ins	Повіває молодою осінню холодна річка з низів.
низів	Case=Gen	Case=Loc	Повіває молодою осінню холодна річка з низів.
закрапало	Aspect=Imp	Aspect=Perf	Із стріх закрапало, а з гір струмочки покотилися.

tagged as feminine by the spaCy parser. Other inconsistencies are found in the part-of-speech annotation.

We intentionally highlight this part as it directly influences the quality of our metrics. Due to the probability of tags' incorrectness, some tokens can be missing from the set; therefore, the final evaluation of the tool may be less precise. At the lexical level, we try to avoid this drawback by checking some explicit morphological characteristics through affixes or the position of a token in the sentence based on a dependency tree. The dependency tags have proven to be the most precise and robust. Hence we tend to rely on them more while implementing grammar and syntactic rules.

C. Metrics assessment

The Ukrainian version of the StyloMetrix incorporates 104 metrics subdivided into lexical forms, parts-of-speech incidence, and syntactic and grammatical structures. The complete list of metrics can be found in Appendix A. In this subsection, we strive to provide general descriptive characteristics and validation criteria for each group.

Table II describes the number of metrics per category. With the StyloMetrix, academics can extract both conventional statistics of the text and features intrinsic to the Ukrainian language. For instance, the universal metrics are the type-token ratio, functional and content words, punctuation, and parts-of-speech statistics. A few examples are presented below.

TABLE III.
TOTAL NUMBER OF METRICS PER GROUP.

Group	Number
Lexical	56 metrics
Grammar	23 metrics
Syntax	14 metrics
Part-Of-Speech	12 metrics

- **L_ADV_POS:** [потрібно, відверто] – positive adverbs [needed, sincerely]
- **L_ANIM_NOUN:** [Президент, агресор, людей] – animated nouns [President, aggressor, people]
- **L_DIRECT_OBJ:** [час, нам, армію, потенціал, альтернативи, режим] – direct object [time, us, army, potential, alternatives, regime]
- **L_INDIRECT_OBJ:** [світом, року, конференції, Україні, режимом] – indirect object (in Ukrainian denoted by case; in English translation we add prepositions) [(by) world, (during) a year, (at) a conference, (to) Ukraine, (in) regime]

Albeit the commonness of these measurements, it has been demonstrated by many researchers, e.g., the Coh-Metrix study, that these scores may provide valuable insight into the idiosyncratic characteristics of a text.

The forms prominent in the Ukrainian language belong to syntactic constructions such as parataxis, ellipses, and positioning (прикладка). Grammatical forms such as two types of the future tense, passive and active participles (дієприсливний доконаного \ недоконаного виду), adverbial perfect \ imperfect participles (дієприкметник доконаного \ недоконаного виду), four types of declensions, and seven cases. For instance:

- **SY_PARATAXIS:** [Я, хотів, чути, від, світу, ", Україна, ,, ми, будемо, з, тобою, "]. – parataxis [I wanted to hear for the world: "Ukraine, we will stand with you".]
- **VF_FIRST_CONJ:** [затримка, підтримкою, помилкою, країна] – first declension [delay, support, mistake, country]
- **L_GEN_CASE:** [виступу, безпеки, лютого, життів, домовленостей] – genitive case (in Ukrainian denoted by suffix) [performance, safety, February, lives, agreements]

The examples are the raw outputs from the metrics, with added translation into English. We evaluate metrics' performance based on the accuracy score assessed by the trained linguist. The best weighted accuracy has been achieved in the part-of-speech metrics – 0.957, due to their reliance on the spaCy tagger. The lexical metrics have obtained a weighted accuracy of - 0.934. Some discrepancies have occurred at rel-

TABLE III.
RESULTS OF THE EXPERIMENTS CONDUCTED IN THE PAPER COMPARED TO THE PAPER BY PANCHENKO ET AL.

Model	**Large training set**	
	Panchenko et al.	This paper
NB-SVM	0.64	-
SM-Voting Classifier	-	**0.66**
Ukr-RoBERTa	0.75	0.82
Ukr-ELECTRA	0.72	0.89

ative and superlative adjectives, adverbs, and case misalignment because of the tagger performance. The grammar group scored 0.912; the inconsistency has occurred in declensions metrics. The syntactic group has got 0.886 in light of the complex constructions, such as parataxis and positioning, that may produce incongruencies.

The accuracy scores indicate that the metrics perform well overall but have some limitations in dealing with complex structures. As the StyloMetrix provides each metric's mean value, a researcher can skip looking into Ukrainian texts to extract the necessary features and conduct further analysis based on the obtained statistics. The descriptions are available for every metric, some with external links to the Universal Dependencies project `https://universaldependencies.org/`.

IV. EXPERIMENTS

This section attempts to represent the StyloMetrix as a baseline for text classification tasks. We further illustrate how to analyze the StyloMetrix baseline model and the possibility of making beneficial inferences about the data based solely on its output.

Conducting a supervised text classification in the Ukrainian language remains challenging due to the scarcity of labeled datasets. There exist a few open-source corpora which can be relevant to this task. For instance, the largest and most popular corpus known by now is UberText 2.0 [5]. The data is subdivided into five smaller datasets: the news dataset, which incorporates short news, longer articles, interviews, opinions, and blogs scraped from 38 news websites; the fiction dataset, with novels, prose, and poetry; the social dataset, covers 264 public telegram channels; the Wikipedia corpus; and the court dataset with decisions of the Supreme Court of Ukraine. The UA news corpus `https://github.com/fido-ai/ua datasets/blob/main/ua_datasets/src/text_classification/README.md` is a collection of over 150 thousand news articles from more than 20 news resources. Dataset samples are divided into five categories: politics, sport, news, business, and technologies. UA-SQuAD is a Ukrainian version of Stanford Question Answering Dataset, and UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian language [27].The list with all state-of-the-art datasets can be found via the link `https://github.com/asivokon/awesome-ukrainian-nlp/blob/master/README.md`.

We ground our experiments on the well-established benchmark public dataset `https://www.kaggle.com/competitions/ukrainian-news-classification/data` provided by Kaggle project. The corpus has been scrapped from seven Ukrainian news websites: BBC News Ukraine, NV (New Voice Ukraine), Ukrainian Pravda, Economic Pravda, European Pravda, Life Pravda, and Unian. Ukrainian computer scientists [23] have developed the described corpus. The researchers give an exhaustive outlook on the data preprocessing steps and the number of texts in the train/test split (57789/ 24765, respectively). The Kaggle platform offers two training splits from the existing sample: large (57460) and small (9299). In their paper, the academics demonstrate their models' performance scores on the two training splits discussed [23]. We are left with the training splits because we cannot use the initial train and test split as it is unavailable to the public.

The large training data partially incorporates the small training sample; hence we leverage the larger corpus, subdividing it into 80/20% training and testing samples, with 15\% for validation. The obtained results are evaluated with macro-averaged F1-score, the same criterion as in the paper. The baseline model leveraged in the study by [23] was Naïve Bayes with SVM; we have added the StyloMetrix with Voting Classifier as our baseline. The Voting Classifier is composed of RandomForest, AdaBoost, and Logistic Regression. As for the main models, we keep the ones utilized in the paper: ukr-RoBERTa [19] and ukr-ELECTRA[26].

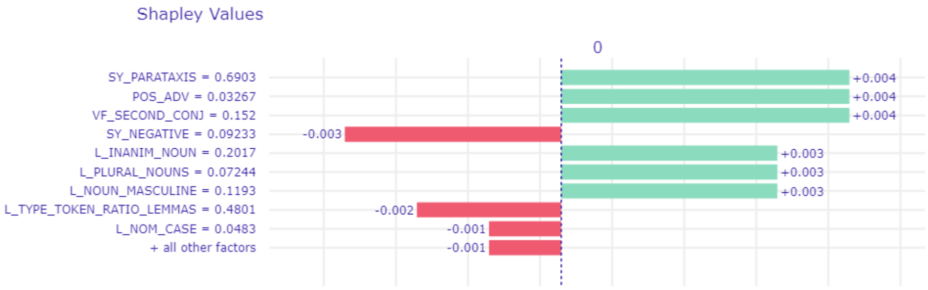
As can be inferred from Table III, the StyloMetrix-Voting Classifier has scored higher than the Naïve Bayes – SVM model but not much, which allows it to serve as a baseline for other more advanced algorithms.

Model Explanation

Unlike the Naïve Bayes – SVM model, the StyloMetrix offers the possibility to extract the descriptive statistics for each group, looking at the most discriminative metrics. For example, we have arbitrary chosen article from class 0 - BBC News Ukraine, to describe the possible data analysis approaches with the StyloMetrix.

One of the wide-used methods of explainable AI is Shapley values [13], which shows the average marginal contributions of features. To make the most of this type of explanation, Shapley values are usually applied to features that can be reversely interpreted, such as categorical values, or to anomaly detection [28]. The most common text representation offering static or dynamic embeddings like GloVe, Word2Vec, or BERT-based vectors does not allow human interpretation of

Fig IVI.
The Shapley values for class 0.



such explanations. The Shapley values indicate the most essential features locally and globally, but the features themselves remain some random columns. With StyloMetrix, on the other hand, the text vector representation consists of interpretable values: each element of the vector translates directly into a given linguistic metric. In this case, indicating the local or global contribution of top features allows for linguistic analysis of grammar or lexical patterns that impact the model's decision when predicting the class.

To implement this, we leverage an open-source library – DALEX [3]. As shown in Fig. II the metrics' significance for class 0 based on their contributions to the model's decision is described. Ultimately, the syntactic metric for parataxis, adverbs, second declension, inanimate nouns, plural nouns, and masculine nouns are prominent in texts that belong to class 0. Vice versa, negative sentences, type-token ratio, and nominative case lower the likelihood of a text falling under this category.

We can dive even deeper into the text statistics and extract the aggregated mean values of the metrics from the StyloMetrix output. As the final vectors are saved in the .csv file, it is easy to find the needed metric and estimate the average mean value for the class. For instance, based on the obtained Shapley, we provide the metrics description and average mean of all texts under class 0 (Table IV). This, in turn, serves the linguistic analysis offering a statistical baseline for a given text genre, including a wide range of metrics. Especially in a multi-class classification, it is vital to compare the baseline against other genres (classes) and draw conclusions about local and global distinctive features.

Therefore, we can conclude that the StyloMetrix as the baseline model can bring some beneficial insights about the texts and the significance of the metrics for a particular classification model. We have presented only one approach to data evaluation with the XAI tool. There are other possibilities to research this area and expand the horizon of the StyloMetrix application and existing metrics.

V. Conclusion

Albeit the idea of constructing the statistical measures of syntactic and grammar features of the text is not new, the experiments discussed in this paper highlight the relevance and significance of creating open-source packages like the StyloMetrix. In the article, we have outlined the main metrics available in the tool's current version and provided some descriptive analysis with the StyloMetrix. We have also discussed the applicability of the corpus analysis tool like StyloMetrix as the baseline "cunny" model for machine learning.

Through experiments, we have traced the metrics importance in a model for classification tasks using the XAI tool – DALEX. More rigorous and detailed analysis is yet to be done in this field, and we consider it the next milestone for our study. The metrics have performed well at the validation step and can be efficient for linguistic analysis of different text genres and the cross-linguistic analysis with other languages such as Polish and English (also available in the StyloMetrix).

References

[1] Bas Aarts, Charles F Meyer, Charles J Alderson, Caroline Clapham, Dianne Wall, and Robert Beard. Livres regus. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 40:3, 1995. Doi: 10.1177/000842987300300410

Table VV.
The StyloMetrix mean values and descriptions of each metric.

Metric	Description	Mean
SY_PARATAXIS	Number of words in sentences with parataxis	0,02731253208958335
POS_ADV	Incidence of adverbs	0,04840493864411134
VF_SECOND_CONJ	Incidence of words in the second declension	0,00027041364427593664
SY_NEGATIVE	Incidence of words in the negative sentences	0,08136554265171815
L_INANIM_NOUNS	Incidence of inanimate nouns	0,013161611405387586
L_PLURAL_NOUNS	Incidence of plural nouns	0,002385407919553026
L_NOUN_MASCULINE	Incidence of masculine nouns	0,20552832194690895
L_TYPE_TOKEN_RATIO_LEMMAS	Type-token ratio for words lemmas	0,054029343395248786
L_NOM_CASE	Incidence of nouns in Nominative case	0,06036709344834804

[2] Karin Aijmer and Bengt Altenberg. English corpus linguistics. *Routledge*, 2014. Doi: 10.4324/9781315845890
[3] Hubert Baniecki,Wojciech Kretowicz, Piotr Piatyszek, JakubWisniewski, and Przemyslaw Biecek. dalex: Responsible machine learning with interactive explainability and fairness in python. *Journal of Machine Learning Research*, 22(214):1‑7, 2021. Doi: 10.48550/arXiv.2012.14406
[4] Douglas Biber. Corpus linguistics and the study of english grammar. Indonesian JELT: Indonesian Journal of English Language Teaching, 1(1):1‑22, 2005. Doi: 10.25170/ijelt.v1i1.93
[5] Dmytro Chaplynskyi. Introducing UberText 2.0: A corpus of modern Ukrainian at scale. In Proceedings of the Second Ukrainian Natural Language Processing *Workshop, Dubrovnik, Croatia,* May 2023. Association for Computational Linguistics.
[6] Rochelle Choenni and Ekaterina Shutova. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. arXiv preprint arXiv:2009.12862, 2020. Doi: 10.48550/arXiv.2009.12862
[7] Noam Chomsky. *Generative grammar. Studies in English linguistics and literature*, 1988.
[8] Peter Collins. It-clefts and wh-clefts: Prosody and pragmatics. *Journal of Pragmatics*, 38(10):1706‑1720, 2006. Doi: 10.1016/j.pragma.2005.03.015
[9] Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with r: a package for computational text analysis. *The R Journal*, 8(1), 2016. Doi: 10.32614/RJ-2016-007
[10] Svitlana Galeshchuk, Arval BNP Paribas, and France Rueil-Malmaison. Abstractive summarization for the ukrainian language: Multi-task learning with hromadske. ua news dataset. *In The Second Ukrainian Natural Language Processing Workshop (UNLP 2023),* page 49, 2023.
[11] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. *Coh-metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers*, 36(2):193–202, 2004. Doi: 10.3758/BF03195564
[12] Sylviane Granger. Automated retrieval of passives from native and learner corpora: precision and recall. *Journal of English Linguistics*, 25(4):365–374, 1997. Doi: 10.1177/007542429702500410
[13] Sergiu Hart. Shapley value. *Springer*, 1989.
[14] Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. Contextual embeddings for ukrainian: A large language model approach to word sense disambiguation. *In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP),* pages 11–19, 2023.
[15] Piaseck Maciej, Walkowiak Tomasz, and Eder Maciej. Open stylometric system websty: Integrated language processing, analysis and visualisation. *Computational Methods in Science and Technology*, 24(1):43–58, 2018. Doi: 10.12921/cmst.2018.0000007
[16] Christian Mair. Quantitative or qualitative corpus analysis? Infinitival complement clauses in the survey of English usage corpus. *Johansson and Stenstrom (eds.)*, pages 67–80, 1991. Doi: 10.1515/9783110865967.67
[17] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press, 2014. Doi: 10.1017/CBO9780511894664
[18] Rahul Mehta andVasudevaVarma. Llm-rmat semeval-2023 task 2: Multilingual complex ner using xlm-roberta. *arXiv preprint arXiv:2305.03300*, 2023. Doi: 10.48550/arXiv.2305.03300
[19] Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics.
[20] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Comput. Surv*, 50(6), nov 2017. Doi: 10.1145/3132039
[21] Inez Okulska and Anna Zawadzka. Styles with benefits. the stylometrix vectors for stylistic and semantic text classification of small-scale datasets and different sample length.
[22] Lingwei Ouyang, Qianxi Lv, and Junying Liang. Coh-metrix model-based automatic assessment of interpreting quality. *Testing and assessment of interpreting: Recent developments in China*, pages 179–200, 2021. Doi: 10.1007/978-981-15-8554-8_9
[23] Dmytro Panchenko, Daniil Maksymenko, Olena Turuta, Mykyta Luzan, Stepan Tytarenko, and Oleksii Turuta. Ukrainian news corpus as text classification benchmark. In *ICTERI 2021 Workshops: ITER, MROL, RMSEBT, TheRMIT, UNLP 2021, Kherson, Ukraine, September 28– October 2, 2021, Proceedings*, pages 550–559. Doi: 10.1007/978-3-031-14841-5_37
[24] Andre Quispesaravia, Walter Perez, Marco Sobrevilla Cabezudo, and Fernando Alva-Manchego. Coh-metrix-esp: A complexity analysis tool for documents written in spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC16), pages 4694–4698, 2016.
[25] Carolina Scarton and Sandra Maria Aluısio. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR*, volume 10. sn, 2010. Doi: 10.1007/978-3-642-16952-6_31
[26] Stefan Schweter. Ukrainian electra model, November 2020.
[27] Oleksiy Syvokon and Olena Nahorna. Ua-gec: Grammatical error correction and fluency corpus for the Ukrainian language, 2021. Doi: 10.48550/arXiv.2103.16997
[28] A Tall'on-Ballesteros and C Chen. Explainable ai: Using shapley value to explain complex anomaly detection ml-based systems. *Machine learning and artificial intelligence*, 332:152, 2020.
[29] Gunnel Tottie. Lexical diffusion in syntactic change: Frequency as a determinant of linguistic conservatism in the development of negation in English. *Historical English syntax*, pages 439–467, 1991. Doi: 10.1515/9783110863314.439

APPENDIX

TABLE V.
PARTS OF SPEECH METRICS

Metric	Description
POS_VERB	Incidence of Verbs
POS_NOUN	Incidence of Nouns
POS_ADJ	Incidence of Adjectives
POS_ADV	Incidence of Adverbs
POS_DET	Incidence of Determiners
POS_INTJ	Incidence of Interjections
POS_CONJ	Incidence of Conjunctions
POS_PART	Incidence of Particles
POS_NUM	Incidence of Numerals

TABLE VI.
LEXICAL METRICS

Metric	Description
L_PRON_RELATIVE	Incidence of relative pronoun 'що'
L_PRON_RFL	Incidence of reflexive pronoun
L_PRON_TOT	Incidence of total pronouns
L_QUALITATIVE_ADJ_SUP	Incidence of qualitative superlative adj
L_QULITATIVE_ADJ_P	Incidence of qualitative adj positive
L_RELATIVE_ADJ	Incidence of relative adj
L_SURNAMES	Incidence of surnames
L_PUNCT	Incidence of punctuation
L_PUNCT_DOT	Incidence of dots
L_PUNCT_COM	Incidence of comma
L_PUNCT_SEMC	Incidence of semicolon
L_PUNCT_COL	Incidence of colon
L_PUNCT_DASH	Incidence of dashes

TABLE VII.
GRAMMAR GROUP

Metric	Description
VF_ROOT_VERB_IMPERFECT	Root verbs and conjunctions in imperfect aspect
VF_ALL_VERB_IMPERFECT	Incidence of all verbs in imperfect aspect
VF_ROOT_VERB_PERFECT	Root verbs and conjunctions in perfect aspect
VF_ALL_VERB_PERFECT	Incidence of all verbs in perfect aspect
VF_PRESENT_IND_IMPERFECT	Incidence of verbs in the present tense, indicative mood, imperfect aspect
VF_PAST_IND_IMPERFECT	Incidence of verbs in the past tense, indicative mood, imperfect aspect
VF_PAST_IND_PERFECT	Incidence of verbs in the past tense, indicative mood, perfect aspect
VF_FUT_IND_PERFECT	Incidence of verbs in the future tense, indicative mood, perfect aspect
VF_FUT_IND_IMPERFECT_SIMPLE	Incidence of verbs in the future tense, indicative mood, imperfect aspect, simple verb form
VF_FUT_IND_COMPLEX	Incidence of verbs in the future tense, indicative mood, complex verb forms
VT_FIRST_CONJ	Incidence of verbs in the first declension
VT_SECOND_CONJ	Incidence of verbs in the second declension
VT_THIRD_CONJ	Incidence of verbs in the third declension
VT_FOURTH_CONJ	Incidence of verbs in the fourth declension
VF_TRANSITIVE	Incidence of transitive verbs
VF_PASSIVE	Incidence of verbs in the passive form
VF_PARTICIPLE_PASSIVE	Incidence of passive participles
VF_PARTICIPLE_ACTIVE	Incidence of active participles
VF_INTRANSITIVE	Incidence of intransitive verbs
VF_INFINITIVE	Incidence of verbs in infinitive
VF_IMPERSONAL_VERBS	Incidence of impersonal verbs
VF_ADV_PRF_PART	Incidence of adverbial perfect participles
VF_ADV_IMPRF_PART	Incidence of adverbial imperfect participles

\

TABLE VIII.
LEXICAL METRICS

Metric	Description
L_DIRECT_ADJ	Incidence of direct adjective
L_QUALITATIVE_ADJ_SUP	Incidence of qualitative superlative adj
L_QUALITATIVE_ADJ_CMP	Incidence of relative adj
L_RELATIVE_ADJ	Incidence of relative adj
L_QULITATIVE_ADJ_P	Incidence of qualitative adj positive
L_ANIM_NOUN	Incidence of animated nouns
L_ADV_CMP	Incidence of comparative adverbs
L_ADV_POS	Incidence of positive adverbs
L_ADV_SUP	Incidence of superlative adverbs
L_DIMINUTIVES	Incidence of diminutives
L_FEMININE_NAMES	Incidence of feminine proper nouns
L_FLAT_MULTIWORD	Incidence of flat multiwords expressions
L_INANIM_NOUN	Incidence of inanimate nouns
L_GIVEN_NAMES	Incidence of given names
L_MASCULINE_NAMES	Incidence of masculine proper nouns
L_NOUN_MASCULINE	Incidence of masculine nouns
L_NOUN_FAMININE	Incidence of feminine nouns
L_NOUN_NEUTRAL	Incidence of neutral nouns
L_NUM_CARD	Incidence of numerals cardinals
L_NUM_ORD	Incidence of numerals ordinals
L_PRON_DEM	Incidence of demonstrative pronouns
L_PRON_INT	Incidence of indexical pronouns
L_PRON_NEG	Incidence of negative pronoun
L_PRON_POS	Incidence of possessive pronoun
L_PRON_PRS	Incidence of personal pronouns
L_PRON_REL	Incidence of relative pronouns
L_TYPE_TOKEN_RATIO_LEMMAS	Type-token ratio for words lemmas
L_CONT_A	Incidence of Content words
L_FUNC_A	Incidence of Function words
L_CONT_T	Incidence of Content words types
L_FUNC_T	Incidence of Function words types
L_PLURAL_NOUNS	Incidence of nouns in plural
L_SINGULAR_NOUNS	Incidence of nouns in singular
L_PROPER_NAME	Incidence of proper names
L_PERSONAL_NAME	Incidence of personal names
L_NOM_CASE	Incidence of nouns in Nominative case
L_GEN_CASE	Incidence of nouns in Genitive case
L_DAT_CASE	Incidence of nouns in Dative case
L_ACC_CASE	Incidence of nouns in Accusative case
L_INS_CASE	Incidence of nouns in Instrumental case
L_LOC_CASE	Incidence of nouns in Locative case
L_VOC_CASE	Incidence of nouns in Vocative case
L_INDIRECT_ADJ	Incidence of indirect adjective

TABLE IX
SYNTACTIC METRICS

Metric	Description
SY_PARATAXIS	Number of words in parataxis sentences
SY_DIRECT_SPEECH	Number of words in direct speech
SY_NEGATIVE	Number of words in negative sentences
SY_NON_FINITE	Number of words in sentences without any verbs
SY_QUOTATIONS	Number of words in sentences with quotation marks
SY_EXCLAMATION	Number of words in exclamatory sentences
SY_QUESTION	Number of words in interrogative sentences
SY_ELLIPSES	Number of words in elliptic sentences
SY_POSITIONING	Number of positionings (прикладка)
SY_CONDITIONAL	Number of words in conditional sentences
SY_IMPERATIVE	Number of words in imperative sentences
SY_AMPLIFIED_SENT	Number of words in amplified sentences
SY_NOUN_PHRASES	Number of noun phrases