Studying Expert-ese: Profiling and Classification of Domain-Specific Language Variation in Architecture with Traditional Machine Learning and LLMs

Carmen Schacht
Ruhr-University Bochum, Germany
Faculty of Philology
Department of Linguistics
carmen.schacht@rub.de

Renate Delucchi Danhier
TU Dortmund University
Department of Cultural Studies
Institute for Diversity Studies
renate.delucchi@tu-dortmund.de

Abstract

This study investigates how domain expertise shapes spontaneous oral language production, with a focus on architecture. Building on the ExpLay Corpus, which contains image descriptions by speakers with and without architectural training, we analyze linguistic variation by combining Profiling-UD and the DECAF framework. We extract a broad range of syntactic and morpho-syntactic features to build linguistic profiles for both groups and train classifiers to distinguish expert from non-expert productions. Two traditional machine learning models (logistic regression and SVM) are compared with a lightweight BiLSTM and two large language models (GliClass and LLaMA 2). While the expert and non-expert corpora diverge only subtly (pairwise Jensen–Shannon divergence (JSD)= 0.25), the BiLSTM using fastText embeddings achieves the highest F1-score (0.88), outperforming both traditional models and LLMs. This indicates that semantic representations are more predictive of domain variation than purely structural features and that smaller neural architectures generalize better on limited data. Overall, the findings provide empirical evidence that architectural expertise leaves measurable linguistic traces in spontaneous speech, supporting the Grammar of Space hypothesis.

1 Introduction

Linguistic variation reflects differences in situational, social, and cognitive contexts. This paper investigates how expertise shapes spontaneous oral language production, focusing on the architectural domain. We operationalize the concept of "expertese": a domain-specific linguistic register analogous to "translationese" (Gellerstam, 1986; Kunilovskaya and Corpas Pastor, 2021; Teich et al., 2016), to describe systematic differences that emerge in the language of expert communities. For the present project, the definition of register adopted is that of Argamon (2019):

> "A register is described by that situational context and the linguistic features typical of the register, along with a description of how those features function specifically within that particular context of language use. That is, the linguistic features describing a register are not arbitrary, but form a complex that is useful for particular communicative purposes in a particular context."

This definition allows for the assumption that linguistic patterns emerge as shaped by their communicative occurrence—in this case, the assumption that domain-expertise shapes the linguistic phenotype of language productions within the respective domain. Previous research on register has shown that linguistic features (lexical, syntactic, and discourse-level) cluster in systematic ways depending on communicative context (Biber and Conrad, 2009). Experts often operate within shared cognitive and informational environments, leading to specialized linguistic patterns (Argamon, 2019; Degaetano-Ortlieb and Piper, 2019). This phenomenon is not limited to writing: spoken language can also exhibit register-specific variation shaped by expertise, although this has been investigated less systematically. Architecture provides a compelling domain for studying these effects, since architects undergo extensive training, use specialized terminology, and engage with spatial representations in ways that differ from laypeople (Mertins et al., 2020). These cognitive and communicative differences are expected to leave measurable linguistic traces even in spontaneous speech. This study builds on insights from information-theoretic approaches to language (Shannon, 1948; Jaeger, 2010), which view communication as the efficient transmission of information. Registers are hypothesized to optimize information flow within communities, often through conventionalized and compact

Proceedings of SIGHUM 2026, pages 16–29
March 28-29, 2026 ©2026 Association for Computational Linguistics

linguistic structures (Halliday, 1988/2004; Teich et al., 2016). If experts share more background knowledge, their speech may exhibit distinctive patterns of linguistic complexity and distribution. Using the ExpLay Corpus (Schacht and Delucchi Danhier, 2025), which contains spontaneous image descriptions by experts and non-experts, we conduct a multi-level analysis of linguistic variation. We apply Profiling-UD and DECAF to extract syntactic and morpho-syntactic features, and evaluate whether machine learning models can distinguish expert from non-expert speech. Methodologically, we explore to which extend the linguistic manifestation of expertise is detectable automatically. Beyond its contribution to register variation research, this work is also motivated by a broader relevance of society: in many public-facing domains, experts have trouble communicating effectively with laypeople. In architecture, where communication with clients, policymakers, citizens, and other interdisciplinary colleagues such as engineers is central to the profession, such misunderstandings can not solely be attributed to a lack of shared specialized architectural vocabulary, but may also reflect differences in how information is structured and distributed in spontaneous speech by experts and laypeople. By showing that architectural expertise leaves measurable traces in oral production beyond professional jargon, this study shifts attention from surface notions of complexity to differences in information packaging and shared knowledge assumptions. While this analysis focuses on architecture, the proposed profiling and classification approach is transferable for investigating expertise-related register variation in other domains where experts and laypeople need to communicate with each other, such as medicine or law. The project data is made available under CC BY 4.0 license[1].

2 Related Work

2.1 Linguistic Complexity and Register Variation

Linguistic complexity has been widely studied in domains such as language acquisition and second language learning (e.g., Park (2024), Lu (2010),Xia et al. (2016), Collins-Thompson (2014),Kyle (2016)). More recently, questions of linguistic complexity have also become prominent in research on machine learning and LLMs, as illustrated in Misra and Mahowald (2024). These strands of research share an interest in identifying and quantifying linguistic features that may serve as indicators of structural or cognitive complexity in language use.

In order to apply linguistic complexity to the analysis of register variation—like Teich et al. (2016) or Kunilovskaya and Corpas Pastor (2021) did—this paper adopts the definition of register introduced earlier, which allows for the assumption of linguistic patterns emerging as shaped by their communicative occurrence. In the context of the present study, this assumption translates into domain-expertise shaping the linguistic characteristics of language productions within a given domain.

This reasoning leads directly to the theoretical frameworks underlying the present study. Following Schacht and Delucchi Danhier (2025), the research builds on perspectives that link domain expertise, cognition, and language production. At its core lies the theory of linguistic relativity (Whorf, 1956; Slobin, 1996), which states that language shapes human cognition and specifically attention. Empirical work has demonstrated language relativity effects in domains including color perception (Winawer et al., 2007; Roberson et al., 2000), spatial reference frames (Levinson, 2003; Majid et al., 2004), and motion events (Papafragou et al., 2008).

Analogous cognitive effects have been observed for expertise in different domains, which can influence perception and processing in ways similar to language. Classic neuroimaging research shows structural adaptations linked to spatial training in taxi drivers (Maguire et al., 2000), while other studies report perceptual and sensory-motor advantages in expert or semi-expert populations such as gamers (Ersin et al., 2022; Jiang et al., 2020). In the architectural domain specifically, eye-tracking studies reveal distinct visual attention patterns for experts compared to laypeople (Delucchi Danhier et al., 2025; Mertins et al., 2020), motivating the extension of such analyses to linguistic behavior.

2.2 Communication Efficiency, Information Theory, and Domain-Specific Conventions

As a continuation of the reasoning outlined above, one of the fundamental principles of communication in general is to achieve a smooth and ideally loss-less transfer of information while eliminating unnecessary linguistic signal—that is, to achieve efficiency—and thereby align flexibly with the communicative situation.

Mathematically speaking, communication al-

[1] https://osf.io/4hvm6/overview.

ways takes place through a noisy channel between the sender (speaker) and the receiver (listener) of information, as described in Shannon (1948)'s Information Theory.

Communication in this sense is understood as a linear transfer of information in single communicative units (for example, in bits) and successful communication as being as loss-less as possible, striving for an ideal use of the communicative channel. This reasoning underlies the framework of the Uniform Information Density hypothesis (UIDh) (Jaeger, 2010), shifting the mathematical-engineering approach towards linguistics. The UIDh proposes that the use of the communicative channel must not exceed its capacity boundaries—neither the upper nor the lower bound—in order to avoid loss of transmitted information. Speakers therefore need to minimize divergence in the flow of information, ensuring a relatively homogeneous distribution of information across sentences.

This is achieved by choosing linguistic encoding in such a way as to keep the flow as constant as possible. From different communicative situations emerge different demands and standards, however, meaning that baseline channel capacities may vary. This leads to the assumption that different registers, for instance, will display distinct patterns of linguistic choices to achieve this balance and further that not only language but also expertise shapes cognition and subsequently linguistic production. According to Teich et al. (2021), this phenomenon manifests in the form of conventionalized linguistic codes serving to smoothen the information density profile of communication in specialized fields while still transmitting the relevant but—compared to general language—heightened amount of information characteristic to the respective field. Concrete examples of this phenomenon are studies on the language in the fields of physics (Halliday, 1988/2004), literary studies (Degaetano-Ortlieb and Piper, 2019), and scientific communication more broadly (Teich et al., 2016; Degaetano-Ortlieb and Teich, 2022), all showing conventionalized and informationally compact linguistic encoding. Similarly, shifts within the same language over time can be registered as shown in diachronic studies of scientific English (Rubino et al., 2016; Degaetano-Ortlieb and Teich, 2018; Biber et al., 2011; Biber and Gray, 2016; Juzek et al., 2020) as well as scientific German (Jakobi et al., 2024). Common structures that facilitate the condensation of information in a production take the form of for example compounding (Degaetano-Ortlieb, 2021; Gamboa et al., 2024, 2025) and metaphor (Halliday, 1988/2004; Webster, 2018), especially in highly technical or scientific language, as they condense a heightened amount of information compared to their phrasal counterparts, further underscoring the link between communicative efficiency, register variation, and domain-specific expertise.

Accordingly, the expectation of denser linguistic encoding as shown by the previous research introduced earlier is well motivated by the UIDh, as shared knowledge among a group of domain-experts can be assumed, lifting the baseline for appropriate information transmission, especially in tasks related to their common professional field where they discuss familiar information. This, in turn, allows for divergence from common structures in general language use and justifies the use of more complex, information-rich expressions, as they will not be considered more complex within their specialized surroundings and thus not spike the flow of information.

2.3 The ExpLay Corpus and Pipeline

Delucchi Danhier et al. (2025) showed that architectural experts and laypeople display distinct visual attention patterns when processing three-dimensional spatial stimuli presented in two dimensions, as revealed by eye-tracking data. In addition to these findings, Schacht and Delucchi Danhier (2025) argue that experts are likely to exhibit comparable differences when verbalizing visual perception, since linguistic production inherently involves the linearization of perceptual input (Levelt, 1989). This perspective integrates principles from Linguistic Relativity, Information Theory, and the UID hypothesis, motivating a systematic investigation of linguistic patterns in expert versus non-expert language.

The study is based on the ExpLay Corpus (Schacht and Delucchi Danhier, 2025)[2], that is available under a CC BY 4.0 license and contains spontaneous German image descriptions produced by speakers with and without architectural training. Participants described a set of architectural stimuli under controlled experimental conditions, resulting in a balanced dataset of expert and non-expert productions. The corpus provides matched elicitation settings and spontaneous language data, making

[2] https://gitlab.ruhr-uni-bochum.de/schaccmr/explay-resource.git.

it well suited for investigating domain-specific linguistic variation.

All data were processed using the ExpLay pipeline, which integrates automatic UD parsing, linguistic feature extraction, as well as profiling via Profiling-UD and DECAF. This infrastructure enables systematic comparisons of linguistic distributions between groups and forms the basis for the analyses presented in this paper.

3 Methodology

Several metrics and automatic analysis tools for the evaluation of linguistic complexity in various use cases exist to date, including the ExpLay resource (Schacht and Delucchi Danhier, 2025) which implements an automatic evaluation module based on the approach of Park (2024), incorporating a selected number of syntactic features automatically extracted from the annotated corpus data.

To extend the approach of ExpLay into an extensive fine-grained analysis, this paper aims to evaluate i) the divergence of the two sub-corpora, quantifying the difference between them and ii) the difference in linguistic profiles of the two sub-corpora by following the objective of Teich et al. (2016), who test if "classes have distinctive linguistic correlates and, if so, how well the classes are distinguished linguistically and which features contribute most to their distinction". Linguistic profiling has received growing attention within the last couple of years and has increasingly influenced research areas such as computational register analysis (Argamon, 2019), which investigates divergences between registers. As Profiling-UD (Brunato et al., 2020) only accepts unparsed data, the raw ExpLay data is re-parsed using the tool, to ensure a consistent analysis throughout this study. The data is then first evaluated using the DECAF framework (Müller-Eberstein et al., 2025), an automatic tool under the MIT license originally designed for the analysis and filtering of annotated datasets to create specialized training-data according to the respective research question or training objective, which also offers a module for the evaluation of divergence among corpora based on their linguistic features. We apply DECAF to the data parsed with Profiling-UD for divergence analysis. Subsequently, the data is processed with the Profiling-UD tool itself, an automatic processing tool freely available online for fine-grained linguistic profiling analysis incorporating more than 130 linguistic features automatically extracted from the data, of which 113 are used in this study, based on their availability in the web-application of Profiling-UD (see the list in F for the sub-set used here). Those were extracted first to identify the features that contribute most strongly to the distinction between the two groups and subsequently to integrate the identified distinguishing features into the classification task, in order to test their predictive strength. The categories of features in Profiling-UD include:

1. Raw Text Properties
2. Lexical Variety
3. Morpho-syntactic Information
4. Syntactic Features

This tool is task-agnostic and explicitly multilingual, as it is built upon the UD-framework (de Marneffe et al., 2021) and thus supports research on language variation flexibly. Based on the profiling results, the current study follows the approach of Schacht and Delucchi Danhier (2025) and Park (2024) by conducting a Principal Component Analysis (PCA) (Jolliffe, 2002) as well as a subsequent ANOVA on the extracted profiles. These steps serve to reduce the multi-dimensional profiling results to the features that contribute most significantly to the variance in the data and to evaluate the statistical significance of these contributions. For both, the PCA and the subsequent creation of the radar charts, the data is normalized applying the standard scalar, mapping the values into a shared dimensional space. The PCA is performed with the Scikit library (Pedregosa et al., 2011) and the ANOVA using the Scipy library (Wes McKinney, 2010), both under 3-Clause BSD license.

The evaluation will focus on the following three research questions derived from the analysis methodology:

1. How much do the two sub-corpora diverge from each other?
2. Do the linguistic profiles of the domain-experts differ from the profiles of the laypeople?
3. Do the domain-experts display more complex linguistic structures than the non-experts?

In addition, two classic machine learning classifiers—a logistic regression model and a Support

Vector Machine (SVM) model—are trained for the classifications of documents into the categories of 'expert' and 'non-expert' following Kunilovskaya and Corpas Pastor (2021), Teich et al. (2016) and Argamon (2019), not to improve classification accuracy, but to validate the linguistic assumption of register-distinction via linguistic features. We first perform feature selection—identifying the most predictive features for our classification task of classifying expert-ese productions—using the grafting technique (Perkins et al., 2003) following Cimino et al. (2017) and subsequently train the classifiers. As these algorithms are feature based algorithms, the classification is used to identify the most predictive features of the two sub-corpora by virtue of the grafting feature selection returning a ranked feature importance list. These features will then be compared to the results from the PCA and ANOVA. The two classifiers are trained using the Scikit library and are set with the default values of the SVM and logistic regression. The grafting is implemented via a custom script and the resulting feature vectors are fed into the models. A five-fold cross-validation is applied to the grafting algorithm. To compare the feature-representation of the data—which can be considered a mostly structural representation capturing linguistic characteristics derivable from the linguistic surface form—with a more semantic approach, which will represent more of the meaning of the input data, a Bidirectional Long Short-Term Memory (BiLSTM) (Hochreiter and Schmidhuber, 1997) model is trained and evaluated on the data as well. For classification, we trained three BiLSTM models using random, GloVe (Pennington et al., 2014) (available under Apache-2.0 license), and fastText (Grave et al., 2018) embeddings (available under CC BY-SA 3.0 license). Models were trained for up to 15 epochs with early stopping and standard hyperparameters. The Scikit library was used for the implementation. In addition, two large language models (the Bert-based encoder model GliClass (Stepanov et al., 2025) available under Apache-2.0 license and the LLaMA-2-7b-hf model from Meta (Touvron et al., 2023), available on HuggingFace upon request under meta license) were evaluated in zero- and few-shot settings to compare the performance of large-scale semantic representations with that of smaller, feature-based architectures. The GliClass model is a variation of the GliNER architecture (Zaratiana et al., 2024) with DeBERTa (He et al., 2020) at its core which has been optimized for classification tasks. We tested the large-v3.0 version of GliClass with 439 Million parameters and the base-v2.0-rac version, which was optimized on a Retrieval-Augmented Classification (RAC) dataset for the few-shot scenario of classification task by augmenting the training with retrieved most-similar examples of the classes which it was trained upon. To run a decoder architecture-designed for generation-against an encoder model, LLaMA 2 7b is also tested. LLaMA-2-7b-hf is a 7 Billion parameter pre-trained, autoregressive decorder language model employing an optimized transformer architecture trained on general language and not fine-tuned on instructions. Tested against the performance of the feature-based models it will compare the predictive power of both dimensions on the divergence of the two sub-corpora. Should only the simpler BiLSTM architecture proof to reliably handle the classification of the two sub-corpora, this might indicate a poor generalization on the part of the LLMs due to insufficient data. In case neither of the models perform well on the task, we can assume the semantic to be non-predictive for the classes.

3.1 Data and Data Collection

The data analyzed in this study originates from the ExpLay Corpus (Schacht and Delucchi Danhier, 2025), which consists of experimental elicitation data in German from spontaneous oral image descriptions produced by participants with and without architectural training and contains 130 productions in total. The corpus has been processed using the accompanying ExpLay-Pipeline. Annotations include morpho-syntactic and lexical parses, dependency parsing within the Universal Dependencies framework (de Marneffe et al., 2021), constituency parsing, as well as compound and coreference annotation. For the corpus, two groups of 13 participants each were selected and matched by gender, age, and multilingual status, to control for potential noise in the data. The expert-group consists of five male and eight female and the laypeople group of four male and nine female participants. The participants' age ranges from 19 and 32 years and both groups include 12 monolingual and 1 bilingual speaker. The degree of expertise in the expert-group has to be categorized as semi-expert though, as the participants were still students of architecture at the time of data collection instead of established architects, and domain-specific effects are still expected accordingly.

4 Results

Calculation of Divergence The calculation of the JSD based on the parses from the Profiling-UD tool resulted in a divergence of 0.25, which is considered a small divergence. Thus, the differences between the two groups might be subtle, but considering the limited amount of data it still proves valuable to evaluate the contribution of features to the divergence. The top 6 features contributing most to the divergence were *Person, xpos, VerbForm, PronType, deprel,* and *upos* (see Table 2 for their strength of contribution).

Even though the contribution of the individual features might be modest, it proves worthwhile to examine how the variation between the corpora takes shape. A closer look into the differences of distribution of the top six contributing features reveals for example variation in the use of person, with the experts using the first person more than the non-experts while the non-experts tend to prefer the third person, indicating a variation in the description of perspective (see Table 3). In addition, experts seem to produce slightly more relative pronouns compared to non-experts (see Table 4), which might indicate a variation in descriptive depth, as relative clauses add additional information to their head. Such variation might point to differences in semantics between the two sub-corpora, motivating the further investigation of linguistic patterns found within the two groups by means of a deeper structural analysis as well as semantic approaches.

Linguistic Profiling of the Sub-Corpora The linguistic profiling was conducted on the unparsed raw data and resulted in a total of 113 features out of the full set of 130 available features automatically selected and used by the parsing tool (see the list in the Appendix F for the subset used in the present study). The features are categorized into the four broader categories presented in Section 3. The syntactic features are then again grouped into subcategories according to Brunato et al. (2020). By normalizing and subsequent pooling of the values of the individual features by group and subcategory, linguistic profiles were created in a shared dimensional space. The standard scaler from the Scikit library was applied for normalization and then the mean was calculated by group per feature. To visualize variations in the different distributions of those profiles radar charts were created, both for by-sub-category and by-broad-category, to compare distributions. As can be derived from the combined category profiles in Figure 1, the greatest variation between the two groups can be observed in the raw text properties. Especially features like token-length or tokens per sentence stand out if we zoom into the radar chart of the raw text properties in Figure 2 in the Appendix B. Experts seem to produce overall longer tokens and longer sentences. Additionally, the type-token-ratio in chunks of 100 tokens in the lexical variety chart (see also Figure 3) is remarkable, as experts seem to produce a greater variety of different tokens than non-experts, indicating more varied semantics. Both of these findings could indicate a more specialized language in domain experts. Also, the morpho-syntactic profiles in Figure 4 in appendix B of the expert sub-group displays more varied characteristics than the profiles of the non-experts, by means of an elevated lexical density and verbal characteristics. With regard to the syntactic features, the profiling seems to be able to replicate the initial findings of Schacht and Delucchi Danhier (2025), as the experts exhibit more pronounced tree structures.

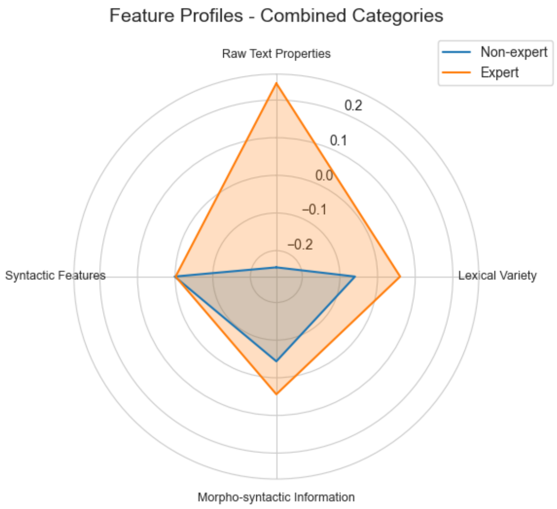

Figure 1: Radar chart of the combined categories profiles.

PCA and ANOVA To also statistically evaluate the visual findings of the linguistic profiling discussed above, a PCA and ANOVA was performed on the feature profiles of both groups. Via the PCA the principle components that reflect the relations among the variables the most can be extracted. By analyzing which variables contribute most to the first principle component (PC1) we can evaluate

Table 1: Classification reports (Micro-averaged Precision, Recall, and F1-scores) of the different classifiers across groups.

Group	**Classifier**	**Precision**	**Recall**	**F1-score**
classic ML	Logistic Regression	0.730	0.730	0.730
classic ML	SVM	0.630	0.620	0.615
BiLSTM	Random Embeddings	0.800	0.650	0.680
BiLSTM	GloVe Embeddings	0.820	0.810	0.810
BiLSTM	fastText Embeddings	0.890	0.880	0.880
GLiClass	large-v3.0/Zero-Shot	0.630	0.530	0.419
GLiClass	large-v3.0/Few-Shot	0.720	0.510	0.350
GLiClass	base-v2.0-rac/Zero-Shot	0.250	0.500	0.333
GLiClass	base-v2.0-rac/Few-Shot	0.750	0.520	0.367
LLaMA2	Zero-Shot	0.740	0.510	0.348
LLaMA2	Few-Shot	0.680	0.630	0.609

the top variables contributing to the variation in the data, as PC1 reflects the greatest explanation of variance. Table 5 shows the explained variation of the first three principal components, with PC1 explaining 15.17 % of the variation. This has to be considered a small contribution, but as in the analysis of the JSD, bound by the limited availability of data, effects are most likely modest.

Following Park (2024), a subsequent ANOVA was performed on the top five contributing variables from the PC1, which can be seen in Table 6. None of them turned out to be significant, but it is still remarkable, that four of them belong to the morpho-syntactic category (auxiliary distributions and upos distributions) and one to the syntactic category (average links length), supporting the finding of Schacht and Delucchi Danhier (2025) of the tendencies to variation in sentence structure, slight as they may be.

Feature Selection and Traditional Machine Learning Classifiers The implemented grafting algorithm iteratively selected six features per classifier according to their predictiveness (see Table 7 for the selected features for the logistic regression model and Table 8 for the features of the SVM). For both models the most predictive feature was character per token, supporting the previously presented importance of this feature. The majority of the remaining features belong to morpho-syntactic distributions, underscoring their predictiveness, as observed in the radar charts already. The classifiers were then trained on the selected features and tested with a held-out test set of 20 percent of the original data set. To compare the performance of all models the micro-averaged F1-scores were calculated for all of the models (see Table 1). Both models performed reasonably well, with the logistic regression outperforming the SVM achieving an F1-score of 0.73 while the SVM only achieved 0.615 (see Table 1). This indicates a relatively strong predictive power of the selected structural features and supports the structural patterns of the respective sub-corpus presented above.

Training of BiLSTM Models and Inference with LLM To test the feature-based classifiers against the model architectures that represent a richer semantic, three different iterations of a BiLSTM were trained using random, GloVe and fastText embeddings. While GloVe captures more global relations, fastText is considered to handle out-of-vocabulary more robustly, which might be relevant in the present scenario, where a specialized domain vocabulary is assumed. Ten models were trained of each iteration and the best performing model was selected. While the model with random embeddings only achieved an F1-score of 0.68—which already outperforms the SVM—the GloVe model achieved 0.81 and the fastText 0.88 (refer to Table 1). This robust performance supports the assumption that semantics might represent the variation in the sub-corpora even more accurately than structural features, as these models are considered to represent the data's semantics due to their context-oriented architecture. Of the LLMs, except for the LLaMA 2 in the few-shot testing, all models performed poorly indicating a poor generalization, probably due to lack of data, as this is a comparatively small dataset. The LLaMa 2 in the few-shot setting achieved an F1-score of 0.609, but the other LLMs all scored below chance (compare Table 1). This suggests that the application of LLMs in research designs with very limited data might not be an ideal choice and the more traditional model architectures might be preferable in these scenarios.

5 Discussion

This study was conducted as a continuation of the initial analysis by Schacht and Delucchi Danhier (2025), extending their approach by comparing the expert and layperson sub-corpora on the basis of linguistic complexity while additionally employing the DECAF and Profiling-UD frameworks to quantify divergence between the groups. This provided a measure of overall corpus-level distance, which was subsequently complemented by a more fine-grained analysis designed to identify the features that distinguish the two groups and to construct domain-specific linguistic profiles. These profiles were then used as input for the training of traditional feature-based machine learning classifiers and compared to the performance of LLMs on the classification task, thereby combining exploratory profiling with predictive modeling.

How much do the two sub-corpora diverge from each other? In terms of quantification of the divergence between the sub-corpora by JSD, the two groups can be said to diverge at a level of 0.25. While this value suggests a small yet measurable difference, the PCA and ANOVA revealed only a small proportion of explainable variation, with no variables contributing significantly. Nonetheless, an inspection of the subtleties of the linguistic profiles revealed tendencies that are in line with the earlier findings of Schacht and Delucchi Danhier (2025), while also pointing towards semantic variation as a promising direction for future research.

Do the linguistic profiles of the domain-experts differ from the profiles of the laypeople? The analysis of the linguistic profiles indicates that the expert group does indeed differ from the non-expert group, albeit subtly. The most prominent differences were found in raw text features, with experts producing longer tokens and longer sentences, as well as displaying greater lexical diversity. Moreover, the structural differences already suggested in the initial analysis of Schacht and Delucchi Danhier (2025) are corroborated by the present findings, reinforcing the assumption of distinct patterns emerging in expert language use.

Do the domain-experts display more complex linguistic structures than the non-experts? The results suggest that the expert group tends to produce slightly more complex phrasal structures compared to non-experts, thus supporting the tendencies observed in the ExpLay study. However, while syntactic complexity plays a role, the predictive characteristics of the two groups appear to be more pronounced in the domain of semantics. While the effects observed in this study are comparatively small, this is not unexpected given that the ExpLay corpus represents only a small corpus of experimental data. In contrast, many studies in register analysis are based on much larger corpora, often spanning millions of tokens, such as the work of Teich et al. (2016) or Kunilovskaya and Corpas Pastor (2021). Considering this, it is not unusual to observe only small effects on limited datasets. What is noteworthy, however, is that even under these conditions the data nevertheless displays tendencies towards register-specific linguistic patterns. Small as they may be, these tendencies indicate that there is indeed detectable variation in the language use of domain-experts within the architectural domain. The differences identified in the present study are subtle, but they mirror the initial findings of Schacht and Delucchi Danhier (2025). In particular, experts display a tendency towards more complex sentence and phrasal structures as well as an elevated use of content words. While some of the structural features proved less predictive in isolation, a closer examination of the individual features reveals that many of these differences ultimately point towards semantic variation. All in all, the combination of several indicators—the divergence in perspective and description depth as reflected in the JSD and the accompanying features, the elevated lexical variety, longer tokens and sentences (suggesting higher content production and potentially greater information density per token, with token length in particular functioning as a proxy for the presence of content words)—all contribute to an overall picture of semantic richness. These findings mirror those of Schacht and Delucchi Danhier (2025), who likewise found an elevated use of content words in the expert subcorpus. The strong performance of the context-based BiLSTM model provides further support for this interpretation. Since BiLSTMs are particularly effective in capturing rich semantic representations of data, their superior performance compared to traditional feature-based classifiers suggests that the predictive characteristics of the corpus are rooted more strongly in semantics than in structural features. Had the classical machine learning algorithms performed comparably or even better, this would have indicated a stronger reliance on structural predictiveness. Instead, the outperformance of BiLSTM with fastText embeddings—well suited to handling out-of-vocabulary tokens that are expected to occur in specialized, domain-specific data such as the present corpus—points clearly towards the presence of predictive semantics. Taken together, these findings closely match the characterization of expert-language as the "use of specialist terminology, nominal style, and high lexical density" of Teich et al. (2016) introduced in Section 1 and support the assumption of Teich et al. (2021) of conventionalized linguistic codes among professionals of a field and those being a cognitive result of expertise emerging from the specific communicative demands in their functional context as Argamon (2019) hypothesized. Thus, while the observed differences remain minor in scope, they nevertheless indicate that domain-specific registers can be distinguished even in relatively small datasets, thereby supporting the claims of capturing representative patterns of domain variation in the line of research of linguistic profiling. Future work could extend this approach by conducting human classification experiments to examine how reliably listeners can distinguish expert from non-expert language. In addition, information-theoretic analyses could explore how differences in linguistic structure relate to patterns of information density across groups. The present study strengthens our understanding of how domain expertise manifests linguistically and opens up future directions to investigate combined structural, semantic and information theoretic characteristics of ‘expert-ese’.

6 Limitations

While this study provides insights into linguistic variation between experts and non-experts, several limitations must be acknowledged. First, the dataset is relatively small, reflecting the difficulty of collecting high-quality, naturalistic elicitation data. This constrains the statistical power of the analysis and the performance of the models. Particularly classifiers typically require substantial amounts of data to be trained and perform robustly. This also affects potential fine-tuning experiments involving pre-trained models, as the limited amount of documents will most likely not be sufficient to fine-tune a pre-trained model on the classification task. The findings of the present study must therefore be interpreted as indicative rather than exhaustive. Second, as the study relies on automatically parsed data, any annotation errors may propagate through subsequent analysis, though state-of-the-art tools were used to minimize this risk. While this issue could be alleviated by more extensive manual curation of the automatic annotations, such curation is very costly in human resources and therefore beyond the scope of this study. Finally, the study focuses exclusively on the domain of architecture, which limits the generalizability of the findings. Replicating the approach across domains and with larger datasets will be essential for testing the robustness and transferability of the observed patterns.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments. This research is partially funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Shlomo Engelson Argamon. 2019. Computational register analysis and synthesis. *ArXiv*, abs/1901.02543.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press.

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *TESOL Quarterly*, 45(1):5–35.

Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Andrea Cimino, Martijn Wieling, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Identifying predictive features for textual genre classification: the key role of syntactic features. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, CEUR Workshop Proceedings. CLiC-it 2017<br/> : Italian Conference on Computational Linguistics ; Conference date: 11-12-2017 Through 13-12-2017.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Stefania Degaetano-Ortlieb. 2021. *Measuring informativity: The rise of compounds as informationally dense structures in 20th century Scientific English*, pages 291–312. John Benjamins Publishing Company.

Stefania Degaetano-Ortlieb and Andrew Piper. 2019. The scientization of literary study. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 18–28, Minneapolis, USA. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific english. *Corpus Ling.. Ling.. Theory*, 18(1):175–207.

Renate Delucchi Danhier, Barbara Mertins, Holger Mertins, and Gerold Schneider. 2025. Entropy as a lens: Exploring visual behavior patterns in architects. *Journal of Eye Movement Research*, 18(5).

A. Ersin, H. Ceren Tezeren, N. Ozunlu Pekyavas, B. Asal, A. Atabey, A. Diri, and İ Gonen. 2022. The relationship between reaction time and gaming time in e-sports players. *Kinesiology*, 54(1):36–42. Doi:10.26582/k.54.1.4.

John Gamboa, Kristina Braun, Juhani Järvikivi, and Shanley E. M. Allen. 2025. The distributional properties of long nominal compounds in scientific articles: an investigation based on the uniform information density hypothesis. *Corpus Linguistics and Linguistic Theory*, 21(1):137–171.

John C. B. Gamboa, Leigh B. Fernandez, and Shanley E. M. Allen. 2024. Investigating the uniform information density hypothesis with complex nominal compounds. *Applied Psycholinguistics*, 45(2):322–367.

M. Gellerstam. 1986. Translationese in swedish novels translated from english. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

M. A. K. Halliday. 1988/2004. On the language of physical science. In Jonathan J. Webster, editor, *The Collected Works of M. A. K. Halliday (Vol. 5)*, pages 140–158. Continuum, London and New York.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. Https://api.semanticscholar.org/CorpusID:219531210.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

T. F. Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Deborah N. Jakobi, Thomas Kern, David R. Reich, Patrick Haller, and Lena A. Jäger. 2024. Potec: A german naturalistic eye-tracking-while-reading corpus. *Preprint*, arXiv:2403.00506.

Chunzhen Jiang, Aritra Kundu, and Mark Claypool. 2020. Game player response times versus task dexterity and decision complexity. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, page 277–281, New York, NY, USA. Association for Computing Machinery.

I. T. Jolliffe. 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York. Doi:10.1007/b98835.

Tom S Juzek, Marie-Pauline Krielke, and Elke Teich. 2020. Exploring diachronic syntactic shifts with dependency length: the case of scientific English. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119, Barcelona, Spain (Online). Association for Computational Linguistics.

S Kullback and R A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

M. Kunilovskaya and G. Corpas Pastor. 2021. Translationese and register variation in english-to-russian professional translation. In V. X. Wang, L. Lim, and

D. Li, editors, *New Perspectives on Corpus Translation Studies. New Frontiers in Translation Studies., . https*. Springer, Singapore. Doi:10.1007/978-981-16-4918-9_6.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University. Doi:10.57709/8501051.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. The MIT Press.

S. C. Levinson. 2003. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge University Press.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

E. A. Maguire, D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. J. Frackowiak, and C. D. Frith. 2000. Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398–4403.

A. Majid, M. Bowerman, S. Kita, D. B. Haun, and S. C. Levinson. 2004. Can language restructure cognition? the case for space. *Trends in Cognitive Sciences*, 8(3):108–114.

H. Mertins, R. Delucchi Danhier, B. Mertins, A. Schulz, and B. Schulz. 2020. The role of expertise in the perception of architectural space. In C. Leopold, C. Robeller, and U. (Hrsg. Weber, editors, *Research Culture in Architecture*, pages 279–288. Birkhäuser, Basel.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Max Müller-Eberstein, Rob Van Der Goot, and Anna Rogers. 2025. DECAF: A dynamically extensible corpus analysis framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 351–362, Vienna, Austria. Association for Computational Linguistics.

A. Papafragou, J. Hulbert, and J. Trueswell. 2008. Does language guide event perception? evidence from eye movements. *Cognition*, 108(1):155–184.

Shinjae Park. 2024. Identifying key linguistic variables of second language speaking proficiency using principal component analysis. *Forum for Linguistic Studies*, 6(6):623–633.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Simon Perkins, Kevin Lacker, and James Theiler. 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. *The Journal of Machine Learning Researchs*, 3:1333–1356.

D. Roberson, I. Davies, and J. Davidoff. 2000. Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3):369–398.

Raphael Rubino, Stefania Degaetano-Ortlieb, Elke Teich, and Josef van Genabith. 2016. Modeling diachronic change in scientific writing with information density. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 750–761, Osaka, Japan. The COLING 2016 Organizing Committee.

Carmen Schacht and Renate Delucchi Danhier. 2025. ExpLay: A new corpus resource for the research on expertise as an influential factor on language production. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 216–227, Vienna, Austria. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

D. I. Slobin. 1996. From “thought and language” to “thinking for speaking". In J. J. Gumperz and S. C. Levinson, editors, *Rethinking linguistic relativity*, pages 70–96. Cambridge University Press.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, Oleksandr Lukashov, Alexander Yavorskyi, and Mykyta Yaroshenko. 2025. Gliclass: Generalist lightweight model for sequence classification tasks. *Preprint*, arXiv:2508.07662.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data-mining approach using register features. *Journal of the Association for Information Science and Technology*, 67(7):1668–1678.

Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. Less is more/more diverse: On the communicative utility of linguistic conventionalization. *Frontiers in Communication*, 5.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jonathan J. Webster. 2018. *18. The Language Of Science – A Systemicfunctional Perspective*, pages 345–363. De Gruyter Mouton, Berlin, Boston.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality*. Cambridge, Ma.

J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.

A K Wong and M You. 1985. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Trans Pattern Anal Mach Intell*, 7(5):599–609.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Divergence Experiments

DECAF operates by running a high-dimensional data analysis on the indexed input data and subsequently calculating the JSD (Wong and You, 1985), which is a metric based on the Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951), which in turn is a quantification of the distance between the distributions of probabilities. Unlike KLD, however, the JSD is a symmetric measure, thus it is not important which of the distributions is being referenced. The divergence was calculated on the the values of the following feature types: *Abbr, Case, Definite, Degree, ExtPos, Foreign, Gender, Mood, NumForm, NumType, Number, Person, Polarity, Poss, PronType, Reflex, Style, Tense, Typo, VerbForm, Voice, deprel, upos,* and *xpos*. Out of the full set, the top six contributing features are displayed in Table 2.

Table 2: Contribution of the top 6 variables in the JSD.

Variable	**Contribution score**
Person	0.070
xpos	0.050
PronType	0.050
VerbForm	0.050
deprel	0.050
upos	0.040

The contribution of the use of person in the two sub-corpora is pointed out in Table 3.

Table 3: Contribution of the use of person in both groups.

Group	**First Person**	**Third Person**
Experts	0.135	0.864
Non-experts	0.071	0.928

Table 4: Contribution of the use of pronoun types in both groups.

Pronoun Type	**Experts**	**Non-experts**
Art	0.625	0.629
Dem	0.033	0.036
Ind	0.144	0.178
Int	0.001	0.002
Neg	0.005	0.005
Prs	0.143	0.105
Rel	0.049	0.041

B Linguistic Profiling Experiments

The following four radar charts display the profiles of the two sub-groups by the four main categories of the linguistic profiling analysis. The first one is depicting the raw text property profiles (see Figure 2) and lexical variety profiles (see Figure 3).

The second radar chart shows the morpho-syntactic information profiles (see Figure 4) and syntactic features profiles (see Figure 5).

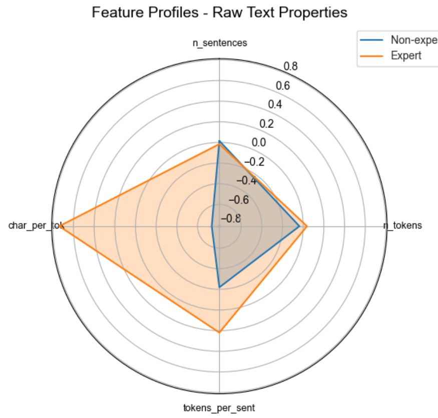

Figure 2: Radar charts of the raw text property profiles.

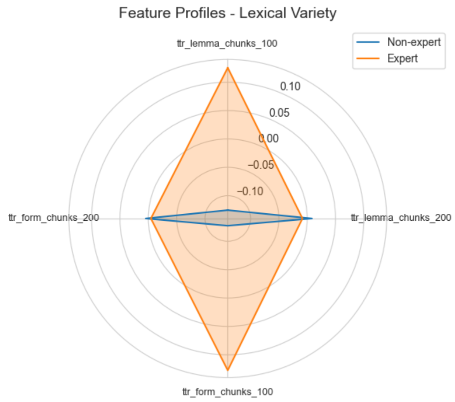

Figure 3: Radar charts of the lexical variety profiles.

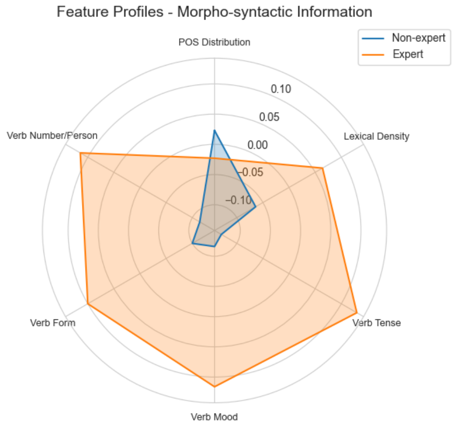

Figure 4: Radar charts of the morpho-syntactic information profiles.

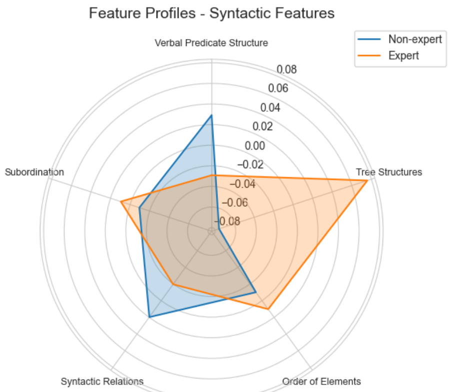

Figure 5: Radar charts of the syntactic features profiles.

C PCA Experiments

The following table show the explained variation of the PCs in the PCA (Table 5 and the results of the one-way ANOVA conducted on the contributing variables from PC1 in Table 6.

Table 5: Explained variation of the components of the PCA.

Components	Explained variation
PC1	15.170%
PC2	13.230%
PC3	9.530%

Table 6: Results of one-way ANOVA for the top 5 contributing linguistic features of PC 1.

Variable	F-statistic	p-value
avg_links_len	1.946	0.176
aux_form_dist_Fin	0.001	0.972
aux_tense_dist_Pres	0.002	0.964
aux_mood_dist_Ind	0.051	0.823
upos_dist_NOUN	0.015	0.904

D Feature Selection via Grafting

Table 7: Grafting feature selection from logistic regression coefficients and their absolute values for the selected features.

Feature	**Abs. Coefficient**
char_per_tok	1.663
verb_edges_dist_0	0.845
dep_dist_root	0.732
dep_dist_csubj	0.502
upos_dist_AUX	0.251
verbs_num_pers_dist_Plur+	0.138

Table 8: Grafting feature selection from SVM coefficients and their absolute values for the selected features.

Feature	**Abs. Coefficient**
char_per_tok	0.491
upos_dist_PART	0.453
dep_dist_csubj	0.173
dep_dist_mark	0.172
ttr_lemma_chunks_200	0.124
verbs_tense_dist_Past	0.029

E Model Evaluation

The following table shows the full classification reports from all tested models.

F Profiling Features

The following list displays the subset of features from the Profiling-UD tool used in the present study: *n_sentences, n_tokens, tokens_per_sent, char_per_tok, ttr_lemma_chunks_100, ttr_lemma_chunks_200, ttr_form_chunks_100, ttr_form_chunks_200, upos_dist_ADJ, upos_dist_ADP, upos_dist_ADV, upos_dist_AUX, upos_dist_CCONJ, upos_dist_DET, upos_dist_NOUN, upos_dist_NUM, upos_dist_PART, upos_dist_PRON, upos_dist_PROPN, upos_dist_PUNCT, upos_dist_SCONJ, upos_dist_VERB, upos_dist_X, lexical_density, verbs_tense_dist_Past, verbs_tense_dist_Pres, verbs_mood_dist_Imp, verbs_mood_dist_Ind, verbs_mood_dist_Sub, verbs_form_dist_Fin, verbs_form_dist_Inf, verbs_form_dist_Part, verbs_num_pers_dist_Plur_, verbs_num_pers_dist_Plur_1, verbs_num_pers_dist_Plur_3, verbs_num_pers_dist_Sing_, verbs_num_pers_dist_Sing_1, verbs_num_pers_dist_Sing_3, aux_tense_dist_Past, aux_tense_dist_Pres, aux_mood_dist_Ind, aux_mood_dist_Sub, aux_form_dist_Fin, aux_form_dist_Inf, aux_num_pers_dist_Plur_3, aux_num_pers_dist_Sing_1, aux_num_pers_dist_Sing_3, verbal_head_per_sent, verbal_root_perc, avg_verb_edges, verb_edges_dist_0, verb_edges_dist_1, verb_edges_dist_2, verb_edges_dist_3, verb_edges_dist_4, verb_edges_dist_5, verb_edges_dist_6, avg_max_depth, avg_token_per_clause, avg_max_links_len, avg_links_len, max_links_len, avg_prepositional_chain_len, n_prepositional_chains, prep_dist_1, prep_dist_2, prep_dist_3, obj_pre, obj_post, subj_pre, subj_post, dep_dist_acl, dep_dist_advcl, dep_dist_advmod, dep_dist_amod, dep_dist_appos, dep_dist_aux, dep_dist_aux:pass, dep_dist_case, dep_dist_cc, dep_dist_ccomp, dep_dist_compound, dep_dist_compound:prt, dep_dist_conj, dep_dist_cop, dep_dist_csubj, dep_dist_dep, dep_dist_det, dep_dist_det:poss, dep_dist_expl, dep_dist_expl:pv, dep_dist_fixed, dep_dist_iobj, dep_dist_mark, dep_dist_nmod, dep_dist_nmod:poss, dep_dist_nsubj, dep_dist_nsubj:pass, dep_dist_nummod, dep_dist_obj, dep_dist_obl, dep_dist_parataxis, dep_dist_punct, dep_dist_root, dep_dist_xcomp, principal_proposition_dist, subordinate_proposition_dist, subordinate_post, subordinate_pre, avg_subordinate_chain_len, subordinate_dist_1, subordinate_dist_2, subordinate_dist_3*