Bialystok Legal Studies
Białostockie Studia Prawnicze
2022 vol. 27 no. 4

sciendo

DOI: 10.15290/bsp.2022.27.04.14

Received: 14.04.2022
Accepted: 21.09.2022

Edyta Więcławska
University of Rzeszów, Poland
edytawieclawska@poczta.fm
ORCID ID: https://orcid.org/0000–0003–0798–1940

Predictive Analysis for Text Classification: Discrete Units in Company Registration Discourse

Abstract: Legal discourse shows variation most commonly in terms of contrasts between languages, textual genres, communicative settings (professional vs. lay communication), translation methods and categories of authors, the last constituting a testing ground for the text-prediction task presented in this article. The research project involves quantitative analysis of selected discrete units and their statistical processing with the R tool for the purpose of generating random forest and decision tree models. It is hypothesised that it is possible to effectively predict text authorship based on the grammatical profile of the texts. The prediction model proposed here covers two authorship categories, institutional name and professional title, and these encapsulate authorship sub-categories related to institutional and work position background. The prediction accuracy parameters for the authorship-based text classification in both cases prove to be statistically satisfactory. More specific findings show that the text classification models for some authorship sub-categories are more effective than for others. Further, some discrete units have distinctively high discriminative power for the texts. The analysis is conducted on a custom-designed corpus, composed of English texts processed in company registration proceedings. The corpus is homogenous in terms of the function and the communicative context of the texts, which assures reliability of the findings and at the same time captures the variationist aspect of legal communication by taking the varied authorship factor into account.

Keywords: authorship factor, decision tree, legal discourse, predictive analysis, random forest, text classification

Introduction

This article fits into the strand of linguistic corpus studies that concern stylistic distinctions in legal discourse[1] based, for example, on the categories of genre,[2] cross-linguistic distinctions,[3] institutional conventions[4] or authorship, the last being largely underrepresented within legal linguistics[5] but extensively present in literary studies as part of stylometric analyses.[6] Here, the authorship-based study involves carrying out predictive analysis, conducted on a corpus of English legal texts, where random forests and decision trees are used for text classification according to context-related variables linked to two authorship categories. The article draws on the concept of variationist linguistics, and specifically it is believed that different drafting styles of distinct authorship categories systematically differentiate legal texts. Two assumptions lie at the root of the task operationalisation, and they are intended to allow the formulation of specific conclusions based on the relevant frequency data. Firstly, it is assumed that the quantitative distinction of a text, noted at the level of specific grammatical categories, translates into the stylistic distinctiveness of various authorship-based text categories. Discrete units, understood as closed class categories, although considered by some to have limited informative capacity,[7] are believed to be of value here as identifying the ground for the syntagmatic research to follow and as providing authentic data on text – prediction; this is legilinguistic research that has not received much attention so far. Secondly, there is no one-to-one correspondence

1 The concept of *discourse* is used here as denoting the nature of the corpus material. It is to emphasise that the material is strongly embedded in the sociolinguistic context and features systemic variation. The inclusion of the authorship-based distinctions is assumed to put the study at the level of discourse level descriptions. The related terms *text* or *language* are used in more specific contexts.

2 V.K. Bhatia, Critical Genre Analysis: Investigating Interdiscursive Performance in Professional Practice, New York 2017.

3 E. Więcławska, Discrete Units as Markers of English: Polish Contrasts in Company Registration Discourse, 'Linguodidactica' 2020, vol. 24, pp. 309–327; E. Więcławska, English/Polish Contrasts in Legal Language from the Usage-based Perspective, (in:) L. Lanthaler, R. Lukenda (eds.), Redefining and Refocusing Translation and Interpreting Studies: Selected Articles from the 3rd International Conference on Translation and Interpreting Studies TRANSLATA III (Innsbruck 2017), Berlin 2020, pp. 99–104.

4 Ł. Biel, Lost in the Eurofog: The Textual Fit of Translated Law, Berlin 2014.

5 The domain of law is represented here by authorship studies conducted for legal purposes (T.D. Grant, Quantitative Evidence for Forensic Authorship Analysis, 'International Journal of Speech Language and the Law' 2007, vol. 14, no. 1, pp. 1–25), rather than by studies conducted on legal texts.

6 D. Longerée, S. Mellet, Towards a Topological Grammar of Genres and Styles: A Way to Combine Paradigmatic Quantitative Analysis with a Syntagmatic Approach, (in:) D. Legallois, T. Charnois, M. Larjavaara (eds.), The Grammar of Genres and Styles: From Discrete to Non-Discrete Units, Berlin 2018, pp. 140–163.

7 *Ibidem*, p. 142.

between the individual authorship categories and textual genres distinguished in the corpus, which ensures that the analysis may lay the ground for identifying new qualitative criteria of text categorisation.

The aim here is to build a prediction model[8] that, above a specific threshold accuracy level, allows for the identification of authors discerned within the two authorship categories of INSTITUTIONAL NAME and PROFESSIONAL TITLE, which are most effectively predicted on the grounds of a distribution scheme of 16 grammatical categories. In order to understand how the texts are assigned to classes corresponding to the authorship categories, and also with the aim of acquiring additional, qualitative knowledge on the subject, two relevant models of decision tree were developed. The statistical calculations fit in the well-established R-tool frameworks, and thus the discussion rests on presenting the results of the quantitative analysis (i.e. relevant models), without focusing on the interim stages of statistical data processing.

The general research question formulated in the study is: Can we construe high-quality prediction models for text classification on the basis of variables related to the authorship factor where the accuracy is above 60%? In other words, does the authorship factor allow us to effectively categorise legal texts automatically? Are the stylistics conventions of legal texts distinct, depending on the categories of the authors?

It is hypothesised that the construed models will have a satisfactory level of accuracy and will enable the identification of statistically significant outcomes. Moreover, the discriminative power of the variables related to the grammatical features and authorship categories varies. Finally, specific patterns exist which are built around a series of consecutive conditions to be fulfilled by the texts that demonstrate repetitiveness and consistency in the grammatical profile of the texts. Further, tendencies are to be discerned regarding the statistical salience of some of these patterns, their frequency and their composition-based scheme (which grammatical categories are involved and what their percentage share is in the prediction models).

More detailed questions are: Does the discriminative power of the individual grammatical features vary, and if so, which grammatical categories have the highest discriminative power in text classification? Further, which authorship-conditioned text categories are most effectively predicted by means of frequency distribution patterns with regard to discrete units? In other words, which authorship sub-categories are the most schematic or emblematic for their stylistics and what are the schemata? And finally, what are the dominating, statistically effective automatic text-identification paths within the prediction models identified?

8 The term *model* relates to the concepts of prediction model or text classification model. *Models* refer to the schemes derived individually for the two authorship categories and/or specific outcomes in the prediction analysis (decision tree, random forest).

1. Methodology

The analysis involved the processing of a custom-designed, monolingual, thematically homogeneous corpus compiled of company registration texts (1, 124, 204 tokens, 932, 839 words). The corpus, referred to as the CorpCourt tool, comprises English legal texts that invariably relate to the same thematic range of company law, and are further limited to the category of texts processed in a court environment for the purpose of company registration. Such an authentic composition of the corpus increases the reliability of the results in that, by capturing the complete range of text types in the said communicative environment and including exhaustive data from court files, it ensures the identification of true and new linguistic distinctions, in our case an authorship-based text classification system that exceeds the classical genre-related text classifications.

The research contributes to linguistic studies on authorship factor which variably take the form of authorship identification/attribution,[9] authorship verification,[10] authorship classification,[11] text categorisation methodologies[12] and plagiarism detec-

9 The terms 'authorship attribution' and 'authorship identification' are considered to be synonymous (E. Stamatatos, A Survey of Modern Authorship Attribution Methods, 'Journal of the American Society for Information Science and Technology' 2009, vol. 60, no. 3, p. 539). M. Bhargava, P. Mehndiratta, K. Asawa, Stylometric Analysis for Authorship Attribution on Twitter, (in:) V. Bhatnagar, S. Srinivasa (eds.), Big Data Analytics. Second International Conference, BDA 2013 Mysore, India, December 2013 Proceedings. New York/Dordrecht/London 2013, pp. 37–47; H. Baayen, H. van Halteren, A. Neijt, F. Tweedie, An Experiment in Authorship Attribution, (in:) Proceedings of JADT 2002, St. Malo 2002, pp. 29–37; H. Baayen, H. van Halteren, F. Tweedie, Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, 'Literary and Linguistic Computing' 1996, vol. 1, no. 13, pp. 121–131; C.E. Chaski, Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations, 'International Journal of Digital Evidence' 2005, vol. 4, no. 1, pp. 1–13; R.M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y-Gómez, P. Rosso, Authorship Attribution Using Words Sequences, (in:) J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, J. Kittler (eds.), Progress in Pattern Recognition, Image Analysis and Applications, New York/Dordrecht/London 2006, pp. 844–853; S. Nirkhi, R.V. Dharaskar, Comparative Study of Authorship Identification Techniques for Cyber Forensic Analysis, 'International Journal of Advanced Computer Science and Applications' 2013, vol. 4, no. 5, pp. 32–35.

10 S. Nirkhi, R.V. Dharaskar, V.M. Thakare, Authorship Verification of Online Messages for Forensic Investigation, 'Procedia Computer Science' 2016, vol. 78, pp. 640–645; H. van Halteren, Author Verification by Linguistic Profiling: An Exploration of the Parameter Space, 'ACM Transactions on Speech and Language Processing' 2007, vol. 4, no. 1, pp. 1–17.

11 S. Kim, H. Kim, T. Weninger, J. Han, H.D. Kim, Authorship Classification: A Discriminative Syntactic Tree Mining Approach, (in:) Proceedings of the ACM SIGIR, July 24–28, Beijing 2011, pp. 455–464.

12 F. Fukumoto, Y. Suzuki, Manipulating Large Corpora for Text Classification, (in:) Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia 2002, pp. 196–203; R. Sprugnoli, S. Tonelli, Novel Event Detection and Classification for Historical Texts, 'Computational Linguistics' 2019, vol. 45, no. 2, pp. 229–265; E. Stamatatos, N. Fakotakis,

tion.[13] This authorship analysis task follows prediction-oriented research,[14] and specifically, it constitutes a testing ground for a text-prediction study, whereby which grammatical features have discriminative power will be investigated and whether and to what extent the frequency distribution scheme of the grammatical features covered by the analysis can constitute a basis for a text-prediction model.

The authorship-related stylometric research is conducted with the application of distinct methodologies, ranging from technologically advanced tools[15] to methodologies more common in literary studies, like cluster analysis.[16] The prediction model presented here is generated with the use of the R-tool methodology.[17]

The methodology applied here, making use of the manually annotated authorship metadata and the text-prediction results obtained in the analysis, contributes to text classification research.[18] The research legitimises the classification of legal texts conducted according to the authorship criterion, which may be considered as complementary to the existing genre-based typologies. Adopting yet another perspective, the methodology employed here contributes to variationist linguistic studies, where distinctions run most commonly through register,[19] type of translation[20] or are determined by institutional conditions. Here the analysis focuses on the distinction criterion that is less commonly studied in the context of legal secondary genres. Legal style varies according to the author, not only according to legal genres as is commonly assumed, and the authorship-based text classification is believed to constitute

G. Kokkinakis, Automatic Text Categorisation in Terms of Genre and Author, 'Computational Linguistics' 2000, vol. 26, no. 4, pp. 471–495.

13 B. Stein, S. Meyer zu Eissen, Intrinsic Plagiarism Analysis with Meta Learning, (in:) Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection, Amsterdam 2007, pp. 45–50.

14 S. Cordeiro, A. Villavicencio, M. Idiart, C. Ramisch, Unsupervised Compositionality Prediction of Nominal Compounds, 'Computational Linguistics' 2019, vol. 45, no. 1, pp. 1–57.

15 E. Stamatatos, A Survey of…, *op. cit*.

16 D. Longerée, S. Mellet, Towards a Topological Grammar…, *op. cit*.

17 N. Levshina, How to Do Linguistics with R. Data Exploration and Statistical Analysis, Amsterdam/Philadelphia 2015.

18 F. Fukumoto, Y. Suzuki, Manipulating Large Corpora…, *op. cit.*; R. Sprugnoli, S. Tonelli, Novel Event Detection…, *op. cit.*; E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic Text Categorisation…, *op. cit*.

19 S. Goźdź-Roszkowski, Patterns in Linguistic Variation in American Legal English, Frankfurt am Main 2011.

20 E. Lapshinova-Koltunski, VARTRA: A Comparable Corpus for Analysis of Translation Variation, (in:) Proceedings of 6th Workshop on Building and Using Comparable Corpora; Association for Computational Linguistics, Sofia 2013, pp. 77–86; E. Lapshinova-Koltunski, Variation in Translation: Evidence from Corpora, (in:) C. Fantinuoli, F. Zanettin (eds.), New Directions in Corpus-based Translation Studies, Berlin 2015, pp. 93–114; E. Lapshinova-Koltunski, M. Zampieri, Linguistic Features of Genre and Method Variation in Translation: A Computational Perspective, (in:) D. Legallois, T. Charnois, M. Larjavaara (eds.), The Grammar…, *op. cit.*, pp. 92–117.

a more general text-categorisation category compared to the genre-based categorisation.

The operationalisation of the task involves generating frequency data for the distribution of specific grammatical categories in texts, as distinct for the authorship categories, and subsequently generating prediction models upon the relevant quantitative data. The discriminative power of specific grammatical categories is assessed based on the quantitative salience of specific values in the random forests and on the prediction potential of the decision trees or parts of these. The operationalisation of the hypothesis is grounded on the potential of the corpus, which rests on the annotation of metadata providing for the authorship information in a representative and comprehensive way. The specificity of the texts (their thematic homogeneity and at the same time their contextual variantivity), together with their authenticity and the involvement of manual data processing at the pre-computational phase, which involved detailed study of the origin of the text, made it possible to identify and record the computational qualification-relevant values referring to authorship categories (hereinafter also referred to as variables) at two levels. Both levels exceed the scope of possible individual stylistic preferences and focus on potential stylistic distinctions emerging from group-specific/collective conventions. Thus, it is expected that the individuals affiliated to one type of institution follow consistent linguistic conventions, and their texts were accordingly annotated with the metadata corresponding to the variable (authorship category) INSTITUTIONAL NAME and to the related variable indicators (authorship sub-categories) labelled ENTITY ENTERED INTO THE REGISTER, AUTHENTIFICATION AUTHORITY, COMPANY REGISTRATION AUTHORITY, COMPANY EXTERNAL ENTITY PROVIDING PROFESSIONAL SERVICES NOT CLASSIFIED ELSEWHERE, and MISCELLANEOUS. The second level of authorship category covered by this analysis was identified according to the same principles; namely, linguistic distinctions are assumed to be noted depending on the work position of the individual drafting a given document, which justified the identification of 13 variable indicators conceptually linked to the variable (authorship category) PROFESSIONAL TITLE. The variable indicators (authorship sub-categories) in question include ENTITY ESTABLISHING THE COMPANY, COMPANY MANAGER, COMPANY OFFICER, ENTITY PARTICIPATING IN THE COMPANY, ENTITY AUTHORISED TO REPRESENTATION, NOTARISATION OFFICER, FOREIGN SERVICE POST, STATE CERTIFICATION AND LEGALISATION AUTHORITY, HEAD OF REGISTRATION AUTHORITY, OFFICER OF REGISTRATION AUTHORITY OF LOWER LEVEL, LEGAL COUNSEL, TAX AUTHORITY, and MISCELLANEOUS.[21] The extract from the database

21 The authorship factor making use of the categories exploited in this study has already been subjected to another analysis conducted by the author, but the analysis was limited in scope with regard to the range of the grammatical categories covered and made use of distinct methodol-

presented below is illustrative of the type of texts making up the corpus and the manual coding system applied. Sensitive data have been removed.

<doc headline="no" paragraph="yes" krs="044" krsitem="3" styear="2008" title="10" professional_title="4" institutional_name="1" country="UK" legal_form="Ltd." stpages="1" stwordcount="2" type_of_translation="1" ttpages="1" ttwordcount="2" sex="K" ttyear="2010">
Company Number: xxx
COMPANIES ACT 1985, 1989 AND 2006 COMPANY LIMITED BY SHARES SHAREHOLDERS' WRITTEN RESOLUTION OF
xxx FIN BET INVEST LTD.
(the 'Company')
The signatories, being at the date hereof the sole members of the Company entitled to receive notice of and to attend and vote at a general meeting of the Company, hereby unanimously RESOLVE and agree the following resolutions pursuant to and in accordance with the Companies Act 1985 (as amended) (the 'Act') and such resolutions shall be for all purposes as valid and effective as if the same had been passed at a general meeting of the Company duly convened and held:
SPECIAL RESOLUTIONS
Share Capital
* IT IS RESOLVED THAT the share capital of the Company of GBP 100 with 100 shares of GBP 1.00 each is hereby sub-divided into 10,000 shares of GBP 0.01 each.
* IT IS RESOLVED THAT the Company's Articles of Association and Memorandum of Association be amended to reflect the change made by the resolution 1 above.
Dated 1 April, 2008 and signed by all members of the Company:
Xxx
Xxx
Xxx
xxx </doc>

As emerges from this extract, the resolution has been assigned the authorship sub-categories of ENTITY ENTERED INTO THE REGISTER (code 1) and ENTITY PARTICIPATING IN THE COMPANY (code 4) under INSTITUTIONAL NAME

ogy (E. Więcławska, Quantitative Distribution of Verbal Structures with Reference to the Authorship Factor in Legal Stylistics, 'Studies in Logic, Grammar and Rhetoric' 2021, vol. 66, no. 79, pp. 147–165). Also, the grammatical categories employed in the foregoing were selectively processed for the identification of generic distinctions, either with regard to the English language alone (E. Więcławska, Sociolinguistic and Grammatical Aspects of English Company Registration Discourse, 'Humanities and Social Sciences' 2019, vol. 26, no. 4, pp. 185–195) or in cross-linguistic perspective (E. Więcławska, Discrete Units…, *op. cit.*; E. Więcławska, English/Polish Contrasts…, *op. cit.*). The present study extends the number of grammatical features in that it provides a cumulative account of the discrete units studied so far and proposes the automatic text-classification methodology of random forests and decision trees.

and PROFESSIONAL TITLE. Coding was conducted on the basis of the text content and, when needed, court file examination.

After the corpus had been manually coded with relevant authorship-related metadata, it was subsequently tokenised and tagged with part-of-speech information according to the standards accepted in related analyses.[22] SketchEngine was used to process and extract the relevant raw frequency data together with the relevant metadata. A random manual check followed the automatic extraction stage.

The construction and interpretation of the prediction models are based on the operationalisation scheme making use of the concept of random forests and decision trees. The analysis involves calculating the frequency-distribution patterns for 16 grammatical categories having the status of discrete units of two types: verbal structures and selected parts of speech categories. These units were selected as grammatical categories that are quantitatively[23] and qualitatively[24] significant for legal stylistics, and, in the case of the verbal structures, the categories covered by the analysis exhausted the repertoire of the verbal structures used in the texts. No other forms were identified in the random sample analysis conducted in the pre-processing stage. The set of verbal structures covered by the analysis includes *modal with past reference followed by active infinitive*, *modal with present reference followed by active infinitive*, *modal with present reference followed by passive infinitive*, *present perfect active form*, *present perfect passive form*, *simple past active form*, *simple past passive form*, *simple present active form* and *simple present passive form*. The remaining grammatical categories involve *adjective*, *noun*, *coordinate conjunction*, *subordinate conjunction*, *adverb*, *pronoun*, and *preposition*.

The data related to the raw frequencies of these categories were extracted from SketchEngine and statistically processed with the R tool to derive the random forest and decision tree models for the two authorship categories. The models are construed on the basis of the normalised data.

The order of the discussion is based on the presentation of the quantitatively overrepresented data, and regarding the random forest data, it involves: (i) identifying the accuracy of the prediction model for the two authorship categories (high quality or not – above 60%); (ii) identifying the variable indicators (authorship sub-categories) that are statistically most prominent within the framework of the two

22 K. Aijmer, Parallel and Comparable Corpora, (in:) A. Lüdeling, M. Kytö (eds.), Corpus Linguistics: An International Handbook, Berlin/New York 2009, pp. 275–291; T. Lehmberg, K. Wörner, Annotation Standards, (in:) A. Lüdeling, M. Kytö (eds.), Corpus Linguistics, *op. cit.*, pp. 484–501; H. Schmidt, Tokenizing and Part-of-speech Tagging, (in:) A. Lüdeling, M. Kytö (eds.), Corpus Linguistics, *op. cit.*, pp. 527–552.

23 M. Gotti, Investigating Specialised Discourse, Bern 2005; C. Williams, Tradition and Change in Legal English, Bern 2005.

24 E. Więcławska, Quantitative Distribution…, *op. cit.*; E. Więcławska, Sociolinguistic and Grammatical Aspects…, *op. cit*.

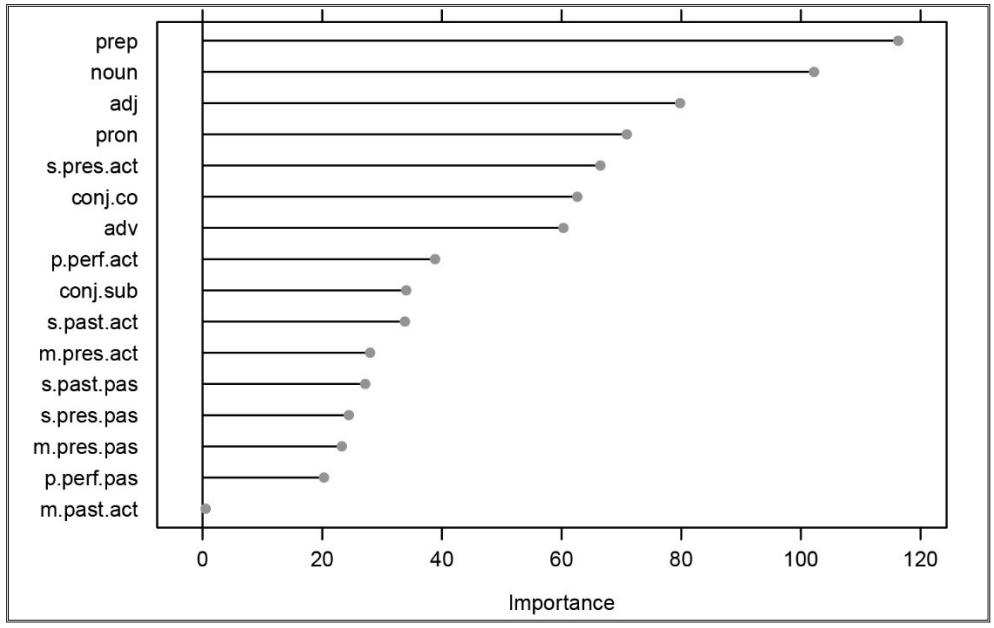
authorship categories; and (iii) identifying the grammatical categories that have the most significant discriminative power in the context of the two authorship categories. In order to obtain the interpretability of the model and extract additional information, decision trees were prepared; the interpretation at this stage is also based on the concept of quantitative overrepresentation of some variable indicators and/or grammatical categories. The trees are scrutinised to identify the most effectively predictable variable indicators (authorship sub-categories) and the linguistic features that are most prominent in terms of their discriminative power for predicting specific texts, that is, texts authored by distinct entities. Closer discussion covers presentation of the prediction path corresponding to one variable indicator that scores the highest percentage value from among those listed in the lowest row of the decision tree (the bottom leaves).

2. Discussion

Institutional Name

The accuracy of the final text-classification model with the variable INSTITUTIONAL NAME is at the level of 87%, which is considered a very good result. Figure 1 visualises the distribution schemes of the 16 grammatical features covered by the analysis, and the interpretation of the data allows us to identify three significance ranges, the borders being assumed at points of significant quantitative divergence between the neighbouring categories.

Figure 1. Discriminative power of the grammatical categories in the prediction model for the variable INSTITUTIONAL NAME

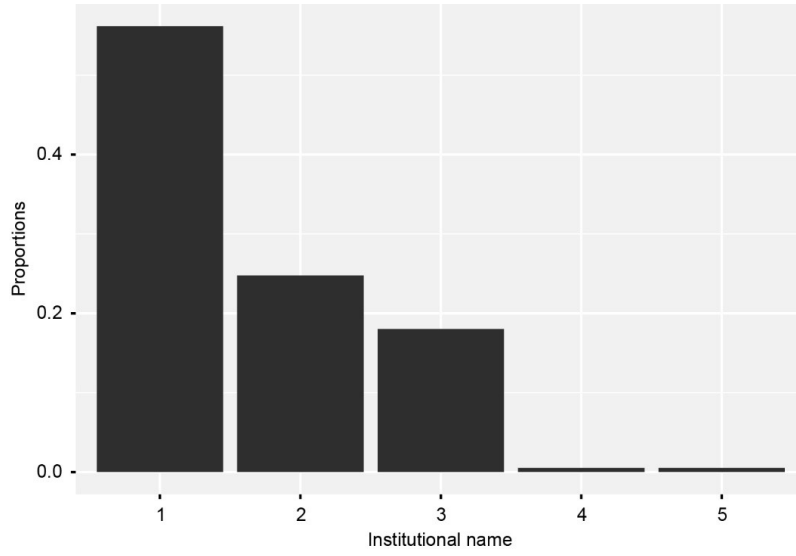


Hence, prepositions and nouns are the leaders among the grammatical categories that have significant discriminative power here, and they may be said to be quantitatively salient in this respect. The next significance range comprises adjective, pronoun, simple present active form and coordinate conjunction, and it closes with adverb. The significance range that is found to be in third place in prominence covers the following grammatical categories: present perfect active form, subordinate conjunction, simple past active form, modal with present reference followed by active infinitive, simple past passive form, simple present passive form, modal with present reference followed by passive infinitive and present perfect passive form. The final significance range is represented by one category, which is modal with past reference followed by active infinitive, which corresponds to almost zero value, confirming the low discursive relevance of this category.

Such distinct distribution of frequency data showing sharp ranges points to (i) varied significance of grammatical categories in the stylistic profile of legal texts, and (ii) marked distinction in the discriminative power between the verbal structures considered as a group and the remaining grammatical part of speech categories.

The random forest model also brings in information related to the types of authors related conceptually to the category of INSTITUTIONAL NAME that are most effectively predicted on the basis of the grammatical categories covered by the analysis. Figure 2 visualises the data in question.

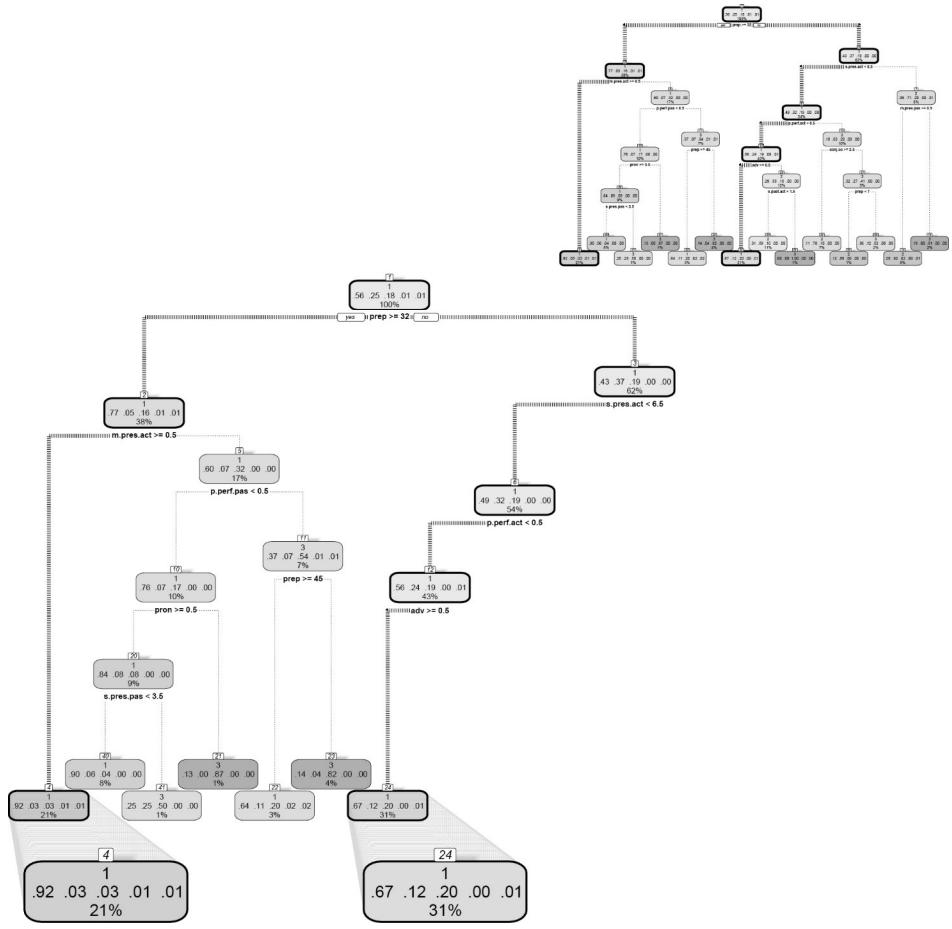
Figure 2. Predictability potential of the variable indicators for the category INSTITUTIONAL NAME



As stated in the previous section, the manual annotation resulted in a 5-degree scale of variable indicators for the category INSTITUTIONAL NAME which were assigned numerical codes as follows: '1' ENTITY ENTERED INTO THE REGISTER, '2' AUTHENTIFICATION AUTHORITY, '3' COMPANY REGISTRATION AUTHORITY, '4' COMPANY EXTERNAL ENTITY PROVIDING PROFESSIONAL SERVICES NOT CLASSIFIED ELSEWHERE, and '5' MISCELLANEOUS. The vertical axis allows us to specify exact values for the variable indicators and single out the most salient ones. Here the relevant values registered in Figure 2 run in descending order, which coincidentally matches the numerical order. As emerges from Figure 2, when it comes to the discriminative power of the said variable indicators, there are two categories. The first three items are shown to be markedly more significant that the other two categories, with low distinction margins among the categories within the two groups. The dominance of the three categories in question can be accounted for by reference to some contextual conditions in which the texts are drafted. Hence, the winning category covers texts authored by entrepreneurs themselves, and the significant homogeneity of the texts and their repetitiveness, and thus high conventionality and prediction potential, is to be attributed to the institutionally recognised stylistic conventions. The same may be assumed to hold true for the authorship category occupying second place. In turn, the shared stylistics and high predictability potential of the texts drafted by the authentication agents is to be attributed to the stylistic conventions imposed on them by the performativity condition. Messages need to be conveyed in a prescribed way, using standard formulae in order to bring about a specific legal effect. The high score of the third topmost variable indicator, that is, COMPANY REGISTRATION AUTHORITY, is to be attributed to the largely prefabricated, form-like type of communication. Here belong, for example, company extracts which are known for their inclusion of tabular-like, automatically generated information.

Based on the statistics used for generating a random forest model, a decision tree was trained with the aim of understanding how specific texts are assigned to the classes. The maximum depth of the tree was kept to 5. The accuracy of the final model is 74%, which is a very good result for a decision tree. Figure 3 is composed of a visualisation of the whole decision tree model placed in the top right corner and a section thereof zooming in on the prediction path corresponding to the variable indicator scoring the highest result with regard to the level of text-classification accuracy.

Figure 3. Decision tree model – INSTITUTIONAL NAME



The decision tree model is discussed by referring to the content of the individual boxes, referred to as leaves on the tree with the numbers assigned to them above (knots) and also with reference to the flow of the prediction paths, the direction of which is conditioned by the fulfilment or non-fulfilment of the conditions specified below. Starting from the topmost row, the information included in the individual boxes (leaves) specifies: (i) the code for the variable indicator corresponding to the authorship category INSTITUTIONAL NAME; (ii) the percentage corresponding to the discriminative power for the said variable indicator in the specific prediction scenario; and (iii) the percentage of the predictability potential with regard to the given prediction path.

In general, the data to be interpreted from the decision tree largely confirm what was stated before, but they also bring in additional information compared to the random forest model and thus enhance the interpretability of the prediction model, specifying the salient prediction paths at the quantitative and qualitative levels. More specifically, the said decision tree model is consistent with the random forest data regarding the discriminative potential of the individual authorship sub-categories discerned within the domain of INSTITUTIONAL NAME. The graphical representation in Figure 3 also reflects the priority order of the authorship sub-categories established according to the values specified in Figure 2. Hence the only three variable indicators here include ENTITY ENTERED INTO THE REGISTER coded with '1', AUTHENTIFICATION AUTHORITY coded with '2' and COMPANY REGISTRATION AUTHORITY coded with '3', represented, for example, in knots 3, 12 and 6 for the code '1' and in knots 13, 26 and 50 and 11, 21 and 23 respectively for the codes '2' and '3'. With regard to the additional information on the text classification potential to be extracted from the decision tree, the data in question (i) disclose the set of grammatical features that are significant for the operation of the salient statistical prediction schemes for the individual variable indicators, and (ii) specify the accuracy level for the prediction scenarios, visualised as prediction paths, and single out the statistically most effective ones. Referring to the first point, in the text classification scenarios (prediction paths) included in the model, almost all the grammatical features are activated in generating the decision tree (11 out of 16). Notably, the missing ones include the top frequency categories, that is, nouns and adjectives. The statistically insignificant participation of nouns and adjectives in the prediction model may be accounted for by the thematic homogeneity of the corpus, which ensures processing of the same concepts/denotations and thus largely the same terms. It is usually nouns and nominal phrases composed of nouns and adjectives that are carriers of legal concepts, and these stay the same throughout the corpus in order to achieve thematic consistency. It may be assumed that the text categories authored by distinctive entities differ in the structures that are relevant for other levels of text organisation. With regard to the second aspect of information to be identified from the decision tree model as complementing the random forest model data, the accuracy level of the prediction paths (text classification) scenarios, as derived from the decision tree model, varies, and in the most general terms, the model accounts for the set of scenarios showing a descending level of accuracy, starting from 11% and ending at 1%. The difference margin among the quantitatively close cases does not exceed 3%. The only exceptions to this pattern are the scores of 21% (knot 4) and 31% (knot 24), as shown in detail in Figure 3, which are markedly salient compared to the others (the difference margin to the closest case is 10%). This testifies to there being two patterns in text-classification models that markedly dominate with regard to the efficiency of the model, both relating to the variable indicator coded as '1', that is, ENTITY ENTERED INTO THE REGISTER.

The logic of the decision tree is described on the basis of a series of conditions that lead to (knot) 24. This leaf/knot, which is the fourth bottom-most leaf counting from the left, contains 31% of the texts from the training group; the dominating variable indicator here is ENTITY ENTERED INTO THE REGISTER, coded as '1'. This authorship sub-category is assigned to 67% of the texts that satisfy the following conditions. The tree shows a series of conditional formulae and the texts are directed to the left or right, depending on whether they fulfil the condition or fail to do so, respectively. The prediction path for knot 24 starts with knot 1, the topmost one, and the prediction scenario is as follows:

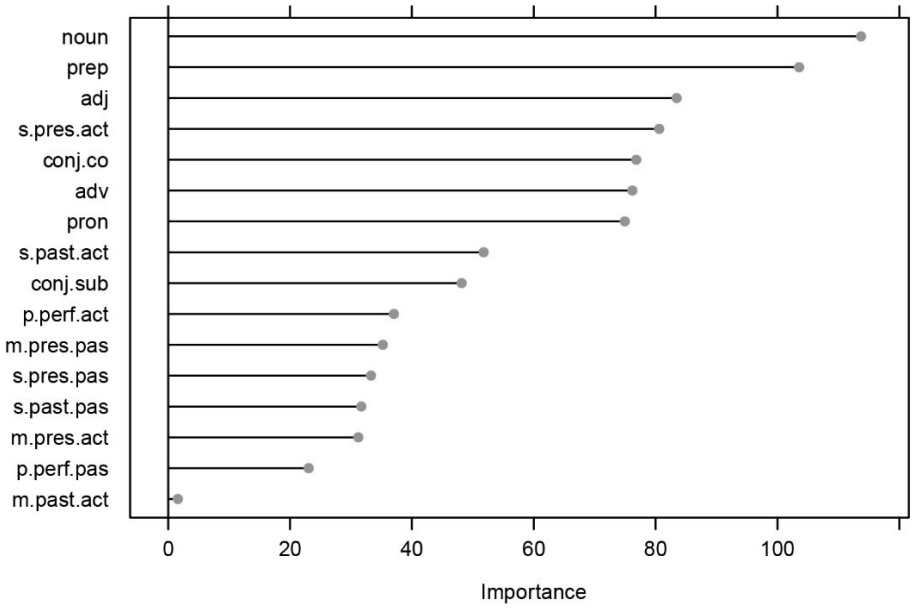
1) 62% of the texts go to the right because they fail to satisfy the condition 'prep>=32',
2) subsequently, 54% of the texts go to the left because they successfully fulfil the condition specified under knot '3', that is 'simple perfect active form <6.5',
3) and then 43% of the texts go to the left since they satisfy the three conditions mentioned so far, including the condition 'present perfect active form <0.5',
4) finally, the last condition is formulated as 'adverb <=0.5'; according to the prediction model 31% of the texts satisfy this condition and the previous ones.

The stylistic structure of the texts is shown to be distinctive by virtue of varied sets of grammatical features, which may be inferred from the range of grammatical categories appearing in the prediction paths generated on the decision tree. To specify, 11 out of 16 grammatical categories act as components of the prediction paths. Furthermore, the participation of verbal structures is significant here. Finally, with regard to the prediction potential of the five variable indicators, a bipolar pattern emerges, where three categories have quantitatively marked text-classification potential and two remain almost at zero level. This can testify to the low stylistic distinctiveness of the texts ranked in the latter group and/or the low level of stylistic repetitiveness, which can hinder the operation of the prediction process.

Professional Title

PROFESSIONAL TITLE is the second authorship category that was identified for verifying the hypotheses posed in this study. The random forest model generated on the basis of this variable noted accuracy at the level of 73%. The importance of the grammatical categories in the model for the authorship category in question is shown in Figure 4.

Figure 4. Discriminative power of the grammatical categories in the prediction model for the variable PROFESSIONAL TITLE



As emerges from Figure 4, the distribution pattern largely resembles the one evidenced for the variable (authorship category) INSTITUTIONAL NAME. The resemblance of the relevant patterns relates to the distribution scheme of the significance ranges, and specifically to the grouping of the data in four ranges, based on the criterion of an insignificant relative distinctiveness margin between the items of the same significance range. Furthermore, the top and bottom grammatical categories are largely the same for INSTITUTIONAL NAME and PROFESSIONAL TITLE. Prepositions and nouns come to the fore here, with the reservation that the order of precedence is reversed compared to the distribution scheme for INSTITUTIONAL NAME. The high position of prepositions in the ranking of their discriminative power can be accounted for by their status as markers of legal discourse in general.25 Nouns prove to rank high in this distribution scheme presumably by virtue of the highly specialised scope of competences ascribed to the authors of the texts, which causes the texts produced to recurrently use the same concepts and thus terms.

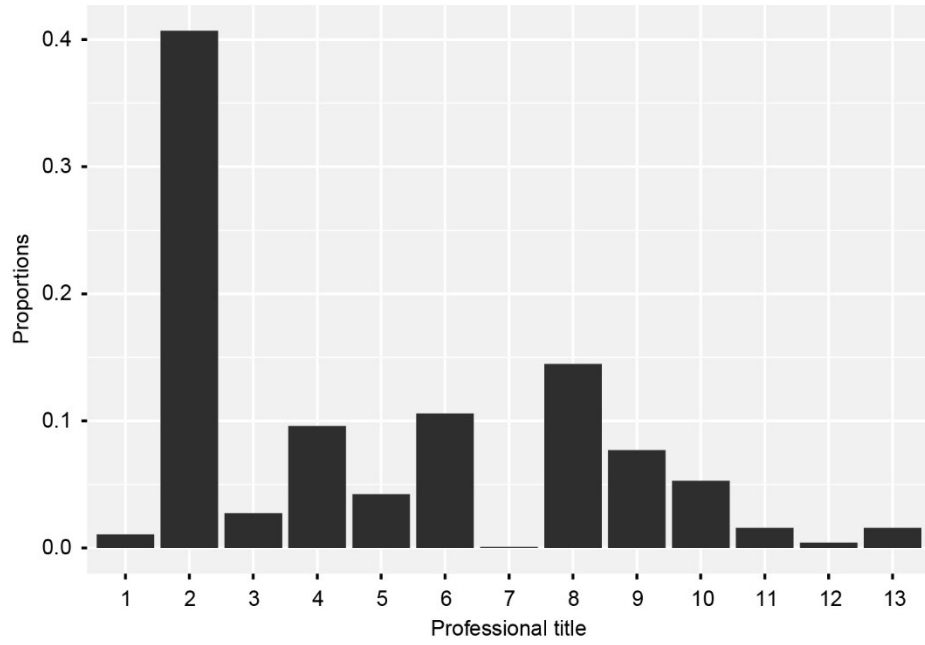
The discriminative power of the grammatical categories with regard to the authorship category PROFESSIONAL TITLE proves to be distinct from INSTITUTIONAL NAME in that the significance ranges corresponding to the middle values

25 Ł. Biel, Phraseological Profiles of Legislative Genres: Complex Prepositions as a Special Case of Legal Phrasemes in EU Law and National Law, 'Fachsprache' 2015, vol. 37, nos. 3–4, pp. 139–160.

are different. Hence, here modal present passive forms and coordinate conjunctions rank lower for PROFESSIONAL TITLE, while simple present forms and modal present active forms score higher values.

Further, the analysis of the random forest model in question provides us with a set of values ascribed to the individual variable indicators and discloses their varied discriminative power.

Figure 5. Predictability potential of the variable indicators for the category PROFESSIONAL TITLE



The horizontal axis in Figure 5 registers individual variable indicators with the numbers corresponding respectively to '1' ENTITY ESTABLISHING THE COMPANY, '2' COMPANY MANAGER, '3' COMPANY OFFICER, '4' ENTITY PARTICIPATING IN THE COMPANY, '5' ENTITY AUTHORISED TO REPRESENTATION, '6' NOTARISATION OFFICER, '7' FOREIGN SERVICE POST, '8' STATE CERTIFICATION AND LEGALISATION AUTHORITY, '9' HEAD OF REGISTRATION AUTHORITY, '10' OFFICER OF REGISTRATION AUTHORITY OF LOWER LEVEL, '11' LEGAL COUNSEL, '12' TAX AUTHORITY, and '13' MISCELLANEOUS.

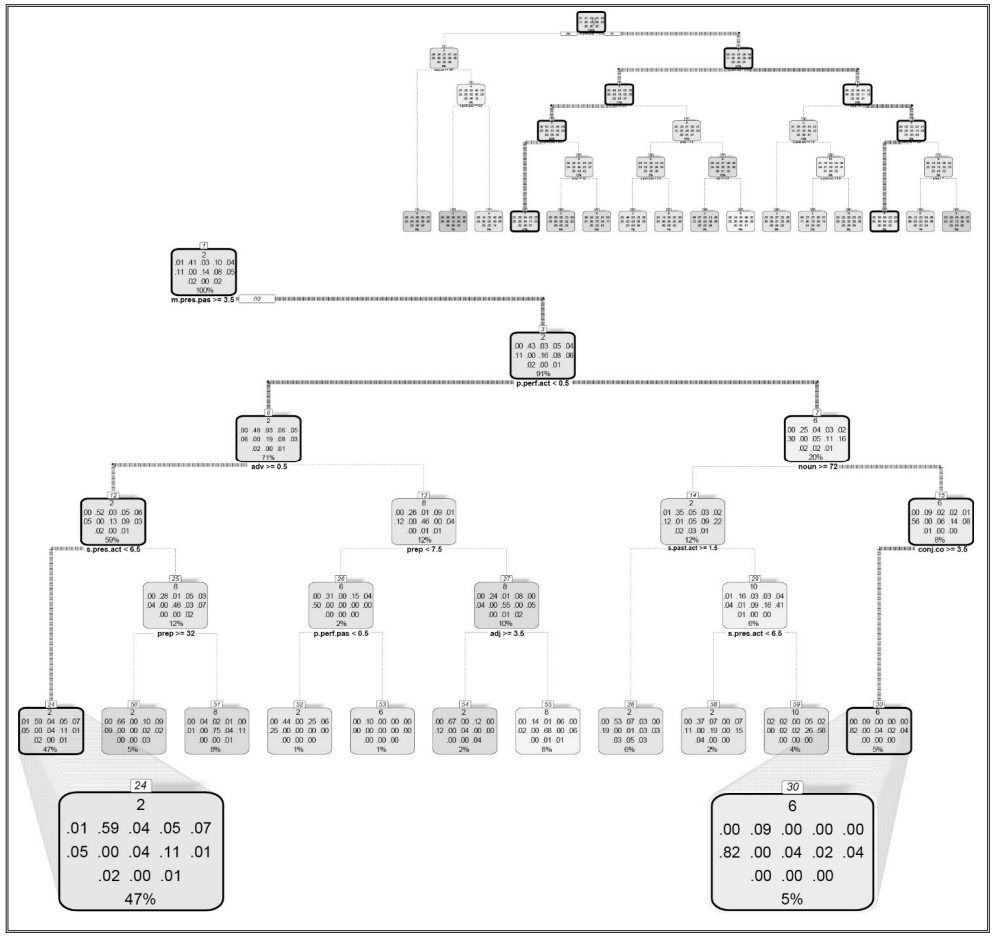
If we interpret the values presented in Figure 5 in relation to the ones registered for INSTITUTIONAL NAME, we see that here the scheme is more dispersed and di-

verse. We have one undisputable winner – COMPANY MANAGER, scoring much higher than the others. The strong stylistic distinctiveness of the texts produced by the entities related to this authorship sub-category may be assumed to be due to there being strong and consistent stylistic conventions for drafting legal documents that are observed by company officers of a higher level and these being distinct from those followed by other agents acting on the professional level, such as, for example, the category NOTARISATION OFFICER. Further, the distribution of the values within the said authorship category can be interpreted by reference to four significance ranges determined by the criterion-of-difference margin not exceeding the value of 5% between the highest and lowest value in the group.

The discriminative force is spread more equally across the authorship category. There are no zero or near-zero values, as was the case for INSTITUTIONAL NAME. The text classification based against these authorship categories is supposed to cover more stylistic details because more sub-categories have been identified at the start and, as emerges from Figure 5, they are shown to have fairly strong discriminative power. The winning category constitutes a significance range of its own with a score of more than 40%. The second significance range covers NOTARISATION OFFICER and STATE CERTIFICATION AND LEGALISATION AUTHORITY. The third group is composed of ENTITY PARTICIPATING IN THE COMPANY, HEAD OF REGISTRATION AUTHORITY, and OFFICER OF REGISTRATION AUTHORITY OF LOWER LEVEL. Finally, ENTITY ESTABLISHING THE COMPANY, COMPANY OFFICER, ENTITY AUTHORISED TO REPRESENTATION, LEGAL COUNSEL, TAX AUTHORITY and MISCELLANEOUS have registered discriminative power at the level of less than 5% and are thus put in the significance range 4.

Plotting the decision tree model adds another dimension to the interpretability data of this prediction analysis. Figure 6 presents the data in question, zooming in on the leaves representing the most effective text-classification models.

Figure 6. Decision tree model – PROFESSIONAL TITLE



The tree here is to be interpreted by reading the technical categories specified for Figure 3. As was the case for INSTITUTIONAL NAME, here the decision tree model (i) confirms the general findings gathered in generating the relevant random forest model, and (ii) allows us to identify the specific text-classification scenarios with regard to the quantitative and compositional context of the individual prediction paths emerging from the model.

Hence the decision tree model confirms the findings emerging from the random forest analysis with regard to the set of authorship sub-categories that have the strongest discriminative power. The three quantitatively topmost sub-categories are COMPANY MANAGER coded as '2', included in knots 1, 3, 12, 14, 24, 50, 52, 54, 28 and 58, STATE CERTIFICATION AND LEGALISATION AUTHORITY coded as

'8', included, for example, in knots 13 and 25, and NOTARISATION OFFICER coded as '6' and included, for instance, in knots 26 and 30.

As before, the decision tree model allows us to complement the random forest model with information regarding (i) the set of grammatical categories that are salient as part of the proposed text-classification scenarios; (ii) the effectiveness result of the prediction model for the individual authorship sub-categories, including the comparative context; and (iii) the compositional structure of the individual prediction paths within the model, including the order of conditions and quantitative conditionings. With regard to the first point, the text classification scenarios composed of the specific prediction paths included in the model exploit the following grammatical categories: simple present active forms, present prefect active forms, present perfect passive forms, and modal with present reference followed by active infinitive, and a set of non-verbal categories that include preposition, adjective, noun and coordinate conjunction. The set emerging here includes the top four frequency categories inferred from the random forest model, and it becomes more selective further down the frequency ladder (Figure 4). For example, when it comes specifically to the verbal structures that invariably scored lower positions in the significance ranking compared to others (Figure 4), simple past active form and modal with present reference followed by passive infinitive are absent from the text classification paths generated as part of the decision tree model.

With regard to the second aspect of information to be identified from the decision tree model as complementing the random forest model data, the accuracy level of the prediction paths derived in the decision tree model is in the range from 47% to 1%. The highest efficiency prediction level is ascribed for the authorship sub-category COMPANY MANAGER, and the logic of the tree is described based on this example. The case in point is knot 24, as shown in Figure 6 and presented against a section of the background data (full graphic in the top right corner of Figure 6) delineated with a somewhat thicker line. This variable indicator of COMPANY MANAGER is shared by 67% of the texts that satisfy the following series of conditions:

1) the number of the modal present passive forms is equal or lower than 3.5 (>=3.5) and thus the prediction path goes to the right,
2) the number of the present perfect active forms is lower than 0.5 (<0.5) and thus the prediction path goes to the left,
3) the number of adverbs is equal to or higher than 0.5 (>=0.5), which directs the prediction path to the left,
4) and finally, the number of simple present active forms is lower than the value 6.5 (<6.5), causing the prediction path to go to the left. In the event that all these conditions are satisfied, the probability that the text was drafted by a COMPANY MANAGER is recorded at the level of 47%.

The findings discussed above allow to us to formulate a few conclusions related to text characterisation. Firstly, the texts of varied authorship within the conceptual domain of PROFESSIONAL TITLE are recognisable by a distinctive grammatical structure, which allows us to generate text-classification models at a satisfactory level of accuracy (57%). Secondly, the stylistic significance of the authorship factor within the prediction scenarios covered by the decision tree for the quantitative and qualitative composition of the grammatical scheme varies, which is confirmed by the distinct discriminative power of the individual authorship categories (variable indicators). Thirdly, the high score of the COMPANY MANAGER coded as the authorship sub-category (variable indicator) '2' testifies to the marked stylistic salience of the texts drafted by the individuals placed in this authorship category; this allows us to conclude that prefabrication and stylistic repetitiveness is a feature of legal texts in general, exceeding the institutional dimension and prescriptive legal texts. Consistency in this sense is noted also with regard to a text drafted in a non-institutional *sensu stricte* environment, exceeding the borders of one country, or one corporation as in our case.

3. Conclusions

It is hoped that this article is a modest contribution to legilinguistic studies approached from the perspective of computational methodology, addressing the issue of intra-disciplinary variation in the grammatical structure of texts produced by distinct categories of authors. It confirms the complexity of legal communication with regard to stylistic conventions and shows that the criterion of genre is not the only one that can be used to classify legal texts. Distinctions and consistency are noted depending on the authorship category. In particular, the author has presented a paradigmatic approach to the automatic detection of a set of grammatical features and has used quantitative data to construe prediction models for automatic text classification where the classes of texts are distinct in that they are produced by different categories of authors.

An additional contribution of this research is that the analysis is conducted on an authentic, custom-designed, manually annotated corpus of texts, representing secondary legal genres. To the author's knowledge, and as voiced in the literature on the subject, such texts are rather understudied in linguistic analyses, and specifically in computational analyses with a focus on text-classification methods. It remains a fact that legilinguistic studies are dominated by institutional (EU) and largely prescriptive texts due to their easier availability and also their larger accessibility for computational processing, thanks to their higher prefabrication level by virtue of institutionally controlled stylistics and ready-made text repositories. Moreover, this analysis includes the context of English as a lingua franca and a global language in legal com-

munication in that the corpus compilation methodology was aimed at making the corpus thematically and situationally homogeneous, and representative for texts of various Anglo-Saxon provenance.

The annotation of the authorship data for the purpose of text classification has been approached in a possibly comprehensive and exhaustive way at the stage of manual annotation. The analysis covered all the relevant data available in the authentic materials, which led to identification of two domains, PROFESSIONAL TITLE and INSTITUTIONAL NAME. The conclusions involve a rough comparison of the values and patterns identified for the two authorship categories, based on the random forest and decision tree models.

The results show that the authorship factor is an effective criterion to classify texts. The most general thesis regarding the accuracy level of the two relevant text-classification models was positively verified, since both models are at a level exceeding 60%. This confirms a few assumptions: firstly, the authorship criterion has significant discriminative power for the texts classified. Secondly, the authorship category can be perceived from the perspective of collective stylistic conventions, not in the traditional individualistic way. Thirdly, the findings show the multi-dimensional character of the authorship variable. The data demonstrate that the prediction models construed for the two distinct authorship categories (INSTITUTIONAL NAME and PROFESSIONAL TITLE) operate effectively and may disclose well-drained details of corpus structure, acting as complementary models.

The secondary theses formulated in this study are also positively confirmed, both by the data emerging from the random forest models and by the decision tree data. Hence, the discriminative power of the individual grammatical features varies, and this model is largely similar for the two authorship categories, with prepositions leading. Further, with regard to the thesis on the presumably varied text-classification potential of the proposed model with respect to the individual sub-categories of authors, the findings point to significant quantitative discrepancies, which leads us to conclude that the stylistic consistency and distinctiveness of the text classes authored by distinct sub-categories of authors vary. Specifically, the texts authored by ENTITY ENTERED INTO THE REGISTER and COMPANY MANAGER are most effectively classified under INSTITUTIONAL NAME and PROFESSIONAL TITLE respectively. Importantly, the consistency and quantitative salience of the patterns identified in the models show that the grammatical structure of the corpus texts remains largely unchanged in the diatopic and diachronic perspectives, the text being varied in terms of publication date and country of origin. This allows us to conclude that repetitiveness and largely schematic stylistics remain a character of legal texts in general, including outside the realm of institutional legal texts, *sensu stricte*, where clear rules and the mostly professional human factor ensure grammatical homogeneity.

The findings and conclusions drawn therefrom deserve further, more detailed, analysis that would extend the qualitative aspect of the data, which at this stage of

analysis was limited to formulating general conclusions, setting the statistics derived in this analysis against the relevant findings gathered in the literature on the subject. It is believed that the analysis of wider lexical context, exceeding the discrete unit paradigm and at the same time including the syntagmatic perspective, would be informative about other linguistic aspects of legal communication. This would bring valuable information regarding the nature and discursive relevance of the grammatical categories that came out here as salient with regard to their discriminative power, and would ultimately allow us to find out whether there are cross-categorial distinctions in the expression of parallel functions with distinct linguistic tools used by distinct sub-categories of authors. Although it was verified at the pre-computational stage that the authorship criterion and genre criterion for text classification do not produce parallel results, it remains to be investigated in more detail how these text-classification models relate to each other, and cross-tabulation analysis would need to be conducted for this purpose. Finally, the proposed model did not take account of the factor of time and place, and the inclusion of the diatopic and diachronic context could bring still finer distinctions with regard to the text-prediction models.

REFERENCES

Aijmer K., Parallel and Comparable Corpora, (in:) A. Lüdeling, M. Kytö (eds.), Corpus Linguistics: An International Handbook, Berlin/New York 2009, pp. 275–291.

Baayen H., van Halteren H., Neijt A., Tweedie E., An Experiment in Authorship Attribution, (in:) Proceedings of JADT 2002, St. Malo 2002, pp. 29–37.

Baayen H., van Halteren H., Tweedie F., Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, 'Literary and Linguistic Computing' 1996, vol. 1, no. 13, pp. 121–131.

Bhargava M., Mehndiratta P., Asawa K., Stylometric Analysis for Authorship Attribution on Twitter, (in:) V. Bhatnagar, S. Srinivasa (eds.), Big Data Analytics. Second International Conference, BDA 2013 Mysore, India, December 2013 Proceedings. New York/Dordrecht/London 2013, pp. 37–47.

Bhatia V.K., Critical Genre Analysis: Investigating Interdiscursive Performance in Professional Practice, New York 2017.

Biel Ł., Lost in the Eurofog: The Textual Fit of Translated Law, Berlin 2014.

Biel Ł., Phraseological Profiles of Legislative Genres: Complex Prepositions as a Special Case of Legal Phrasemes in EU Law and National Law, 'Fachsprache' 2015, vol. 37, no. 3–4, pp. 139–160.

Chaski C.E., Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations, 'International Journal of Digital Evidence' 2005, vol. 4, no. 1, pp. 1–13.

Cordeiro S., Villavicencio A., Idiart M., Ramisch C., Unsupervised Compositionality Prediction of Nominal Compounds, 'Computational Linguistics' 2019, vol. 45, no. 1, pp. 1–57.

Coyotl-Morales R.M., Villaseñor-Pineda L., Montes-y-Gómez M., Rosso P., Authorship Attribution Using Words Sequences, (in:) J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, J. Kittler (eds.), Pro-

gress in Pattern Recognition, Image Analysis and Applications, New York/Dordrecht/London 2006, pp. 844–853.

Fukumoto F., Suzuki Y., Manipulating Large Corpora for Text Classification, (in:) Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia 2002, pp. 196–203.

Gotti M., Investigating Specialised Discourse, Bern 2005.

Goźdź-Roszkowski S., Patterns in Linguistic Variation in American Legal English, Frankfurt am Main 2011.

Grant T.D., Quantitative Evidence for Forensic Authorship Analysis, 'International Journal of Speech Language and the Law' 2007, vol. 14, no. 1, pp. 1–25.

Halteren H. van, Author Verification by Linguistic Profiling: An Exploration of the Parameter Space, 'ACM Transactions on Speech and Language Processing' 2007, vol. 4, no. 1, pp. 1–17.

Kim S., Kim H., Weninger T., Han J., Kim H.D., Authorship Classification: A Discriminative Syntactic Tree Mining Approach, (in:) Proceedings of the ACM SIGIR, July 24–28, Beijing 2011, pp. 455–464.

Lapshinova-Koltunski E., Variation in Translation: Evidence from Corpora, (in:) C. Fantinuoli, F. Zanettin (eds.), New Directions in Corpus-based Translation Studies, Berlin 2015, pp. 93–114.

Lapshinova-Koltunski E., VARTRA: A Comparable Corpus for Analysis of Translation Variation, (in:) Proceedings of 6th Workshop on Building and Using Comparable Corpora. Association for Computational Linguistics, Sofia 2013, pp. 77–86.

Lapshinova-Koltunski E., Zampieri M., Linguistic Features of Genre and Method Variation in Translation: A Computational Perspective, (in:) D. Legallois, T. Charnois, M. Larjavaara (eds.), The Grammar of Genres and Styles: From Discrete to Non-Discrete Units, Berlin 2018, pp. 92–117.

Lehmberg T., Wörner K., Annotation Standards, (in:) A. Lüdeling, M. Kytö (eds.), Corpus Linguistics: An International Handbook, Berlin/New York 2009, pp. 484–501.

Levshina N., How to Do Linguistics with R. Data Exploration and Statistical Analysis, Amsterdam/Philadelphia 2015.

Longerée D., Mellet S., Towards a Topological Grammar of Genres and Styles: A Way to Combine Paradigmatic Quantitative Analysis with a Syntagmatic Approach, (in:) D. Legallois, T. Charnois, M. Larjavaara (eds.), The Grammar of Genres and Styles: From Discrete to Non-Discrete Units, Berlin 2018, pp. 140–163.

Nirkhi S., Dharaskar R.V., Comparative Study of Authorship Identification Techniques for Cyber Forensic Analysis, 'International Journal of Advanced Computer Science and Applications' 2013, vol. 4, no. 5, pp. 32–35.

Nirkhi S., Dharaskar R.V., Thakare V.M., Authorship Verification of Online Messages for Forensic Investigation, 'Procedia Computer Science' 2016, vol. 78, pp. 640–645.

Schmidt H., Tokenizing and Part-of-speech Tagging, (in:) A. Lüdeling, M. Kytö (eds.), Corpus Linguistics: An International Handbook, Berlin/New York 2009, pp. 527–552.

Sprugnoli R., Tonelli S., Novel Event Detection and Classification for Historical Texts, 'Computational Linguistics' 2019, vol. 45, no. 2, pp. 229–265.

Stamatatos E., A Survey of Modern Authorship Attribution Methods, 'Journal of the American Society for Information Science and Technology' 2009, vol. 60, no. 3, pp. 538–556.

Stamatatos E., Fakotakis N., Kokkinakis G., Automatic Text Categorisation in Terms of Genre and Author, 'Computational Linguistics' 2000, vol. 26, no. 4, pp. 471–495.

Stein B., Meyer zu Eissen S., Intrinsic Plagiarism Analysis with Meta Learning, (in:) Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection, Amsterdam 2007, pp. 45–50.

Więcławska E., Discrete Units as Markers of English: Polish Contrasts in Company Registration Discourse. 'Linguodidactica' 2020, vol. 24, pp. 309–327.

Więcławska E., English/Polish Contrasts in Legal Language from the Usage-based Perspective, (in:) L. Lanthaler, R. Lukenda (eds.), Redefining and Refocusing Translation and Interpreting Studies: Selected Articles from the 3rd International Conference on Translation and Interpreting Studies TRANSLATA III (Innsbruck 2017), Berlin 2020, pp. 99–104.

Więcławska E., Quantitative Distribution of Verbal Structures with Reference to the Authorship Factor in Legal Stylistics, 'Studies in Logic, Grammar and Rhetoric' 2021, vol. 66, no. 79, pp. 147–165.

Więcławska E., Sociolinguistic and Grammatical Aspects of English Company Registration Discourse, 'Humanities and Social Sciences' 2019, vol. 26, no. 4, pp. 185–195.

Williams C., Tradition and Change in Legal English, Bern 2005.