DOI: 10.23817/strans.16-2

Received: 10.03.2025
Accepted: 03.06.2025

STUDIA TRANSLATORICA
2025 • vol. 16 33–47
ISSN 2084–3321 e-ISSN 2657-4802

Edyta Więcławska
University of Rzeszów/ Poland

Phraseology at the service of automatic text classification: Computational account of binomials in the company registration discourse with translational implications

Abstract

Phraseology at the service of automatic text classification: Computational account of binomials in the company registration discourse with translational implications

The paper problematises text classification potential of phraseology in the context of methodological research paradigms of digital humanities. Specifically, the analysis discussed in the paper aims at verifying whether four structural features of English legal binomials can construct a statistically reliable genre predictive model, whereby legal texts are automatically assigned to three groups of legal genres, processed through company registration, in a court-based procedure. The computational part of the analysis rests in generating a random decision forest. The candidate terms have been extracted from a corpus of authentic texts, referred to as company registration discourse. The quantitatively most marked patterns are subjected to qualitative analysis involving the examination of the linguistic context, the contextual metadata and the genre requirements. The analytical part is preceded by a theoretical account of the concept of digital humanities with reference to related notions and a historical overview of how this discipline evolved. The findings related to the predictive capacity of the model derived are discussed in the context of the translation output, the translation quality and the comprehensibility of legal texts. At this stage the discussion draws on the parallel component of the corpus. The author concludes that legal genres covered by the analysis are relatively schematic and the structural profile of binomials makes up part of some of these schemes. Automatic text classification is possible in some cases. The original generic structure of the texts covered by the analysis is distorted in the translation.

Keywords: genre prediction, decision tree, predictive model, legal phraseology, digital humanities, computational stylometry

1. Introduction

The analysis problematises the linguistic issues that may be said to belong to the field of digital humanities, if we use this term in its extensive definition, that is the 'application of computing to cultural heritage'. The aim of the paper is twofold. The practical part concerns the presentation of a probabilistic model of text classification (objective 1). In order to set clear ground for the practical part the author presents the findings emerging from the theoretical analysis, whereby the recent developments in the field of digital humanities are described in the context of synchronic and diachronic factors, such as overlapping terminology, evolution in time and status-quo assumptions (objective 2). It is hypothesized that the terminological vagueness with regard to the language computing techniques calls for their systemisation and explanation by reference to data emerging from critical analysis of the literature of the subject (hypothesis 1). With regard to the first objective set to this analysis, the author formulates a hypothesis concerning the possibility of generating a statistically reliable predictive model for the text classification in a company registration communicative environment (hypothesis 2).

2. Methodology

The theoretical part of the paper exploits the critical analyses of the research papers that are claimed to belong to the digital humanities *sensu largo*. The methodology at this stage involves the comparison of the methodological approaches and theoretical paradigms applied over the years with the specific focus on the most contemporary studies (hypothesis 1). The methodology employed for the verification of hypothesis 2 (analytical part) focuses on computing the frequency of specific structural types of binomials with the aim of deriving a statistically reliable predictive model for the classification of texts for their genre affiliation. The model is based on the distinctions of various structural types of binomials (variables and values). 8550 binomials, defined as 'same-word class, coordination, most often linked by a conjunction' (Więcławska 2023a) were extracted from a corpus of texts qualified as company registration discourse[1] (Więcławska 2021; 2022; 2023a; 2023b). The candidate

1| The corpus is composed of texts processed by legal actors and submitted to court files of the National Court Register (KRS) with the purpose of registration of a branch of foreign company. Two divisions of the KRS were searched and all the texts on files as of 2017 were digitalised and processed. The term 'discourse' (cf. company registration discourse) here is used to signal the rich contextual background used for annotation of the digitalised corpus and considered in the qualitative analyses.

terms were extracted with the use of the Sketch Engine tool and then processed by the R tool.

This methodology was already used to conduct predictive analysis on the corpus data covered by this paper, but the previous analyses set different specific goals (Więcławska 2021). This model accounts for an extended predictive scheme, where we deal with three categories of genres, each comprising a few contextually and linguistically related genres, manually assigned to the specific categories (groups).

The translation-related component of this analysis is based on the examination of the parallel material in the Polish language with the aim to derive findings on the translation shifts and relate them to the concept of genre integrity in Polish and English.

This analysis relates to other studies that may be considered to be methodologically close to prediction analyses. The variable of authorship distinction is a case in point (Bhargava et al. 2013; Coyotl-Morales et al. 2006; Nirkhi/Dharaskar 2013).

3. Discussion

3.1. Defining the whereabouts for digital humanities: The reference point for this analysis

Digital Humanities (DH) is a term that covers a range of frequency-based research paradigms, proposed over the last 70 years (cf. quantitative turn).[2] Researching of this issue follows the research trends of frequency-based linguistic investigations (Creswell 2013; Fabiszak/Konat 2013; McIntyre/Walker 2019; Mellinger 2017). The 'quantitative turn' shifted the interest of researchers to frequency data and frequency-based argumentation. This enabled conclusions to be drawn on the basis of usage-based data and on the basis of narrower sets of data that were contextually homogeneous. Various community-related patterns and variations were taken into account. In order to understand the concept of DH and set up a more up-to-date theoretical background for the analytical part, reference needs to be made to how capacious the notion of DH is. Focus will also be on some research trends associated with the notion of 'quantitative turn' in linguistic studies and on the disciplines that emerged,

2| The label DH as a name of a discipline is reported to have been proposed around 2006 (Nyhan/Flinn 2016: 1), but the related concept has been in operation since the 1950s, when Busa started compiling concordance hits that were extracted from some literary Latin works (Busa 1980 as cited in Nyhan/Flinn 2016: 1). Before 2006, a time limit having been proposed somewhat arbitrarily, different terminological labels were in use, that is *Humanist Informatics*, *Humanities Computing*.

either preceding the emergence of the term DH or arising as secondary terms denoting more closely delineated disciplines or sub-disciplines. The terminological richness in this respect is the result of the evolution of frequency-based linguistic research. The concept of DH is treated here as an overarching term denoting the linguistic research approach that this analysis fits into with regard to the research objectives set, framework and methodology employed.[3]

DH has been assigned many definitions that vary on the ground of the time-related, methodological and research-object related perspectives adopted by scholars. Conceptualisation of this term is also found to be problematic when we attempt to find foreign language equivalents for it (Gogora 2020). In this synthetic account we shall follow the general-to-more-detailed account, the latter discussed in the context of its overlap with either its historical quasi-synonyms or with the related terms denoting various fragmentary areas of the DH research approach. Hence, according to the most general definitions, DH is '[a domain] …handling with digitalised versions of knowledge products' (Maci/Sala 2022: 1) or '[…] application of computing to cultural heritage […]', '[…] intersection of computing and cultural heritage' (Nyhan/Flinn 2016: 1). In doing so, DH is referred to as a 'macro-domain', 'linguistic discipline' (Maci/Sala 2022) or '[research] approach', these denotations being indicative of the relatively independent status of DH.

Some scholars claim the beginnings of DH go back to c. 1949 (Nyhan/Flinn 2016: 1) when the first signs of computing being used in literary research were identified. Others see the source of development of DH in specific theoretical shifts, leading to the usage-based approach to linguistic research, this being associated with the publication of Langacker's work (1987). The advent of the usage-based approach (cf. 'quantitative turn in linguistics' (Levshina 2015: 2)) consists in the basic assumption that language competence is not an idealist, mentalist model, but it is shaped by usage, the latter being conditioned by a number of context-related factors, with an outreach to the issue of language variation, language diversity and other issues in the field of interest of sociolinguistics (language used by minorities, language used over specific time periods by specific professional groups, in language communities). Further, exploiting computationally-enhanced, quantitative methodologies, which was possible with new technological resources available, enabled researchers to pursue novel aspects of the tradition of cognitive studies, that is identification of novel, onomiasiologically linked conceptual paths. Quantitative analyses based on automatic data extraction, non-supervised extraction brought the

3| Cf. the semantics of the other terms related to DH, that is 'computational stylistics' and 'computational stylometry' is discussed in Więcławska (2025).

researchers to the new level of analysis by inclusion of new data in the linguistic analysis. Advanced extraction procedures (unsupervised extraction) gave scholars the possibility of identifying patterns not legitimised on the ground of Structuralism and/or Generativism, but which could be accounted for by reference to cognitivist mechanisms, where individual semantic fields are extended by items embodying specific metaphorically or metonymically derived secondary senses and/or items departing from the prototypical semantic profiles (Geeraerts 1983; 1997). The interest shifted from what was mainstream language competence, legitimised on the ground of grammar books, to what is often marginal in terms of frequency but also in terms of how structurally conventional it is.

Finally, the technologically-conditioned focus on the research of usage-based language patterns brought about developments in the field of phraseology, where distributional phraseological paradigms and the unorthodox, structurally more loose, unconventional items started to be treated as phraseological units (Grabowski 2015; Granger/Paguot 2008; Pęzik 2013).

To sum up, although the definitions of DH referred to the abovementioned were coined with the use of the word 'computing' and 'digitalisation', the spread of DH is not to be seen as exclusively due to the development of technology. Technology made it possible to address issues that were not on the research agenda before, but DH developed out of the rise of natural interest in what is beyond 'innate linguistic competence' (Levshina 2015), but remains a fact, and it is a linguistic performance determined by various context-specific factors (Geeraerts 2010, as cited in Levshina 2015: 3). DH was often presented as a novel research paradigm, standing for the second concept in the following dichotomic pairs: 'idealist theories' vs. 'rigorous empirical methodology', 'innate linguistic competence' vs. '[focus on] frequencies or probabilities' (Levshina 2015: 2). The usage-based research paradigm, as DH was labelled, was also defined as 'recontextualisation' (Geeraerts 2010, as cited in Levshina 2015), which is indicative of the tendency to look for linguistic patterns that are less quantitively marked, but significant in terms of their discourse functions (Freake et al. 2011).

In order to understand well the concept of DH, reference needs to be made to related terms which are either quasi-synonymic, overlapping, conceptually related by a metonymic relationship, or they have different referents but are mutually related on the ground of processing digitalised texts with the use of computational methodologies. Most traditionally DH is discussed in the context of its relation to the concept of corpus linguistics (CL). The common denominator for these concepts is that they both involve 'data retrieval and investigation' or – if we relate to the digital aspect of the related processes and the statistical element – 'handling digitalised texts' and 'computational approaches'

(Maci/Sala 2022: 2). Here we should not mistake the phrase 'computational approaches' with the phrase 'computational stylistics', the first being used in reference to a range of statistical methods applied, in a way that is not commissive to any research domain using statistical methods.

The mutual interdependence of the concepts DH and CL, that is the most general terms related to the research domains exploiting computational methods, is not always expressively specified. The somewhat vague conceptual load of these terms is due to the fact that over time scholars have proposed a number of narrow definitions addressing specific aspects of this concept. DH was variously defined as a set of methodologies, research domain, field of investigation or domain of practice (Maci/Sala 2022). In this comparative framework, following most recent theoretical literature surveys DH is assigned the status of a superordinate concept ('super-ordinate term', 'macro-research area') and its distinctiveness is verbalised with words conceptually related to the semantic domain of quantity. In comparison with CL, DH means: (i) a more extensive choice of texts submitted to analysis, (ii) less structured corpora that the analyses are based on, (iii) a wider range of objectives set to specific research tasks and (iv) more IT tools (Maci/Sala 2022). More specifically, DH concerns analyses covering literary texts, unlike CL, which prefers specialised texts. Further, DH, unlike CL, also involves analyses conducted on 'unstructured' datasets. CL covers analyses that are conducted on purpose-built corpora, compiled according to the criteria of a specific environment, including exact annotations, while the analyses couched in the DH paradigm are also conducted on unstructured data, raw data; the Clarin infrastructure may serve as an example here (Does et al. 2017; Kučera 2022; Wissik et al. 2022; Vaičenonienė et al. 2024). Further on, with regard to the range of objectives set, DH may be said to embrace more advanced handling of digitalised texts (e.g. text mapping, transposition). While CL deals mainly with the quantification[4] of candidate terms and inferring specific language patterns on the basis of frequency scores, DH is more quality-oriented. In the latter, the methodology involves not only identification but also localisation of specific linguistic elements and the focus lies on ascribing them specific discursive, stylistics, rhetorical and pragmatic functions. The opposition of 'identifying and attesting patterns of language' vs. 'identifying and locating […] elements' corresponds to the opposition between CL and DH respectively (Maci/Sala 2022: 3). The objectives of the research that may be said to fit in the DH research

4| Reservation needs to be made with regard to the hypothesis about CL being more quantity-oriented. This is meant to be a generalisation and we need to acknowledge the importance of human intervention when processing digitalised texts (Maci/Sala 2022).

paradigm range from sociolinguistic variation (Biber/Finegan 1994), contrastive linguistics issues (Neumann 2014), automatic text classification methods (Fang/Cao 2015) to genre evolution (Moretti 2013).

With such status quo, DH is referred to as 'super-ordinate term', 'macro-research area' ('super' and 'macro' signalling conceptually wider concepts), while CL has been designated with the label of 'language-based discipline' or 'domain of practice in data retrieval and processing' (Maci/Sala 2022: 1–4).

With the development and sophistication of the computational methods used to study literary texts, there appeared new terms coined to precisely define narrower scopes of computational approaches that were informative about specific research objectives (Frontini et al. 2018: 118–119). This trend is represented by the tradition of stylometry, also referred to as computational stylometry by Oakes (2009: 18), a discipline that focuses mainly on the authorship-related issues (cf. authorship attribution) (Frontini et al. 2018: 119). Other scholars claim that stylometry also deals with the issues of genre and text identification (Legalloiset al. 2018: 167), and the term 'identification' here is to signal the opposition to the term 'linguistics of characterization', the latter signalling qualitative studies. For some scholars the realm of text and genre identification belongs to the domain of *computational stylistics* (Frontini et al. 2018: 119).

3.2. Predictive analysis for text classification in the commercial law environment

The practical part of this paper concerns the predictive analysis that follows the research paradigm requirements of DH in the sense discussed above[5]. The analysis has been conducted on the basis of digitalised material, the conclusions have been drawn on the basis of a computationally derived model and the methodology goes beyond basic quantitative argumentation.

Following the first part, the operationalisation of the research task related to hypothesis 2 consisted in the statistical processing of 8550 candidate terms (binomials), annotated for the variables corresponding to the structural profile. The variable of number of conjuncts accounted for the distinction into two-element sequences and longer ones (NOC/code 'a' and code 'b'). The variable of semantic motivation distinguished between binomials proper, that is synonymous sequences and those where items were conjoined by relation of complementarity in four variants (MOT/codes 'synonyms', 'antonyms' 'grammatical pairs' and

5| Previous research on predictive models conducted on comparable language data (Więcławska 2024) adopted a narrower typological perspective and the related analysis was claimed to fall to the realm of computational stylistics (Frontini et al. 2018: 119).

'sequential patterns'). Further, binomials were found to structurally differ on the ground of the linking element between the conjuncts (LEL/codes 'a', 'b' and 'c' corresponding to the use of 'and', 'or' or a comma or a combination of 'and' and 'or'). Finally, the part of speech distinction accounted for the values of nominal, verbal, adverbial, adjectival and prepositional binomials (POS/codes 'nominal', 'verbal', 'adverbial', 'adjectival', 'prepositional'). The data extracted from Sketch Engine were statistically processed with the R tool to derive decision trees (also a predictive model) for the three categories of genres (group 1, group 2 and group 3). Groups 1 covers authorisations, authentications and verifications in the amount of 517. Group 2 encompasses certificates of registration and company extracts in the amount of 527. Group 3 is composed of resolutions, reports and declarations of will in the amount of 523 (cf. notary public environment vs. court environment vs. in-company environment). The three bolded predictive paths correspond to the statistically most effective ones for the three groups of genres.

The interpretation of this predictive model follows the principles of the tree structure logic. We start from the top-most position that can be compared to a tree trunk and we follow along the individual branches illustrating the set of consecutive conditions to be met in order to proceed to the next location on the tree. The bottom-most boxes, also referred to as tree leaves to relate to the

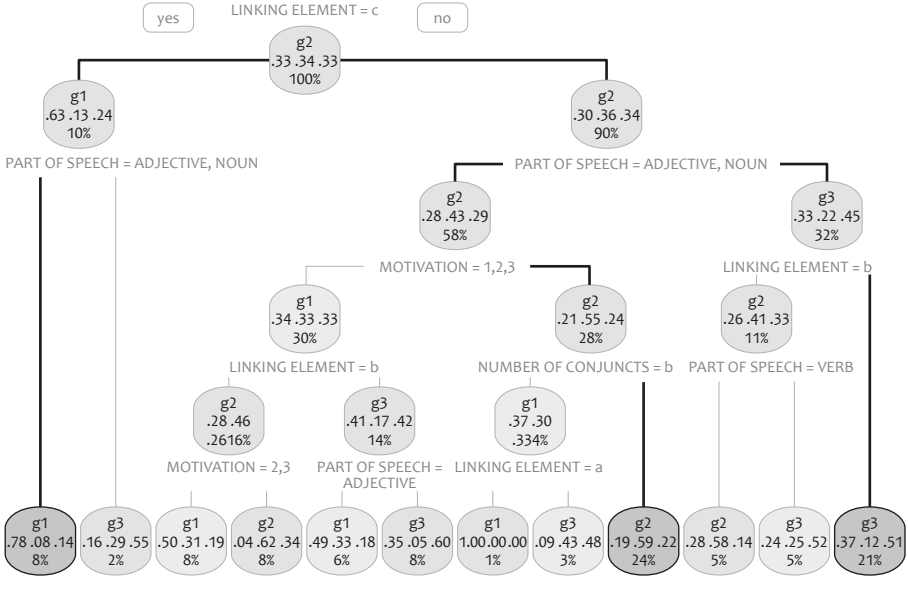

Figure 1: Prediction tree for text classification: The case of notary public environment vs. court environment vs. in-company environment distinction (compiled by Więcławska)

parallel plantosemic nomenclature, include information about the efficiency of the predictive model.

In the bottom leaves, the scores can be found from which the scope of prediction capacity of the model is inferred. Three of the bottom-most leaves are bolded and these were selected for verbal explication since they are found to be represented by the statistically most effective predictive paths in the model. The predictive model derived for the data corresponding to the three groups of genres confirms varied prediction potential for the three groups of genres. Specifically, the statistical efficiency of text classification for group 2 and group 3 is comparable, with the status at the level of 24% and 21% respectively. Both of these predictive paths start with singling out the population of binomials, whereby the conjuncts (elements of binomials) are conjoined by 'and' (LEL/code 'a' included) or 'or' (LEL/code 'b' included and LEL/code 'c' eliminated). For group 3 the next conditions to be fulfilled in the consecutive order involve the inclusion of prepositional, verbal and adverbial binomials (POS/ codes 'adjective' and 'noun' eliminated) and then removal from the population of the obtained sequences, linked by 'or' (LEL/code 'b' eliminated). As noted above, the first condition of the computational text classification model for group 2 is the same as for group 3. The next three conditions making the predictive path for group 2 involve consecutively: (i) selecting adjectival and nominal binomials) (POS/code 'adjective' and 'noun' selected), and then (ii) selecting sequential patterns only (MOT/code '4' included), and finally (iii) eliminating sequences that are longer than two-element sequences (NOC/code 'b' eliminated).

The prediction path for group 1 is operational at the lower level of statistical efficiency (8% of the total number of candidate terms fall into this category). These texts make up the population that is obtained when we choose binomials, where conjuncts are conjoined by comma or combination of 'and' and 'or' (LEL/code 'c' included) and out of the population of texts obtained in such a way we filter out those that are not nominal or adjectival binomials (POS/ code 'adjective' and 'noun' included).

The individual predictive paths in the model are obviously to be looked at from the position of a holistic (prediction) mechanism embracing a set of mutually related conditions, related by the rule of precedence and a yes/no qualification with regard to satisfaction of the consecutive conditions visualised as a tree. Importantly, success, that is reaching the prediction efficiency at the levels specified in the bottom-most leaves, consists in the gradual curbing of the population of candidate terms to those that satisfy (go to the left on the model) or do not satisfy (go to the right) the consecutive conditions related to the structural profile of candidate terms. 'Gradual' here means that the selection of candidate terms occurred at consecutive stages.

The illustrative material below brings to the fore those structural features that are found to markedly construe this predictive model at its individual levels, that is they materialize most of the conditions of the relevant predictive path for a specific group. With regard to the part of speech variable, the adjectival and nominal binomials are found to have significant discriminative power, in the case of texts affiliated to a court environment (group 2), notary public environment (group 3). Further, the linking element parameter is another variable that plays a role in the three most effective predictive paths, as emerged from the analysis. The semantic motivation and the number of conjuncts show as a variable having lower discriminative power and they only make up part of the predictive model for texts affiliated to a court environment (group 2).The most representative nominal binomials include those that are conceptually related to validity (*force and effect*, *tenor and validity*) or authentication (*accuracy or completeness*; *duties*, *powers or office*; *duties or powers*; *irregularity or invalidity*; *power*, *authority or discretion*; *power and authority*; *power or authority*; *power or capability*; *powers*, *authorities and discretions*; *powers and discretions*; *powers and duties*; *powers or rights*; *seal or stamp*; *signature*, *seal or stamp*).[6] In the in-company environment adjectival binomials are predominantly related to legalisation procedures. The main pattern emerging from qualitative analysis shows that for the notary public environment most binomials belong to the conceptual category of verification, credibility (e.g. *true and correct*) and authorisation (e.g. *power and authority*).

The results of the predictive analysis may also be discussed from the perspective of genre congruity with regard to the parallel data in Polish, in the context of the efficacy of English language predictive models for the Polish parallel material. The quantitative analysis of the relevant translation patterns in the population of two element sequences (Więcławska 2023b) confirms significant translational shifts, which – after qualitative analysis of related random candidate terms – allows us to conclude about significant loss of genre congruity. The extensive operation of the simplification processes in its three manifestations[7], i.e. reduction, insertion and conjunction distortion (2,294 hits for reduction/53.15107%, 2,131 hits for insertion/49.37442% and – 3,001 for

6| For more on this see Więcławska (2023b).

7| Reduction is deemed operative whenever binomials are translated as one word (e.g. '*by or through* agents […]' / '*poprzez* agentów […]'. Insertion consists in destabilisation of coordinate structure by separating two components of a binomial with an object or otherwise (e.g. '*unless and to the extent* that the same shall be impossible […]'/ '*o ile będzie to możliwe i w stopniu w jakim to będzie* możliwe […]'. Conjunction distortion destabilises the semantics and structural frame of the source text candidate terms by employment of non-dictionary equivalents for conjunctions, and thus weakening idiomaticity (e.g. '*collaterally with or to the exclusion of* […]'/ '*czy to równolegle z, czy z wyłączeniem* […]').

conjunction distortion/69.53197%) distorts the generic profile of the texts and – as emerges from the qualitative analysis of random, sample data – weakens the predictive potential of the Polish material.[8]

The analysis of the parallel data in question has been found to confirm the findings formulated with regard to other categories of genres (Więcławska 2025) in that Polish equivalents often represent cases of conceptual load loss, since translations are shown to prioritise comprehensibility parameters (Kubacki/ Więcławska 2022). The negative assessment of the translational shifts affecting binomials (cf. 'conceptual loss', 'distortion') is grounded in the fact that there is no consistent approach being followed. Case studies conducted on various populations of contemporary legal texts confirm random and non-standardised decisions taken by translators where no account of specific context is taken (e.g. Dobrić Basaneže 2018), while the varied motivation of binomials has been ascertained by many scholars (Graën/Martin 2021). This conclusion has also been confirmed by the reception study conducted on the corpus exploited in this paper (Kubacki/Więcławska 2022). The study shows that the relevant comprehensibility model favoured by professionals, as opposed to lay users is to be largely based on legal contrastive analysis.

The translation patterns have been analysed for genre dependency. No significant interdependencies have been noted. The translation shifts affect equally all the three groups of genres, and they are found to lead to generic distortion and incapacitation of genre-predictive models throughout the discourse examined.

4. Conclusions

The theoretical part of this analysis is an attempt at systematizing the terminology connected with computational approaches. It is noted that the related terms do not arise in an ad hoc manner as denotations of the same concepts, but they are rather the result of the evolution of new disciplines and new studies being published in the field of computational approaches. We see that the exploitation of computational approaches is a field where many hypotheses are tested and an increasing number of new tools are being deployed. Further, ever new types of texts (not only literary texts) are being subjected to computational

8| The relevant statistics on translation patterns identified in the company registration discourse were published and discussed in the context of their scope and translation variation factors (Więcławska 2023b). These findings have been subjected to further analysis in the context of genre integrity and text prediction-related implication with distinct categories of genres incapsulated in the same corpus (cf. Więcławska, submitted for publication).

analyses, and all these factors set the ground for identifying ever new macro- or sub-disciplines, if we treat the concept of DH as the most general term in point. We note that the relevant concepts are linked to specific, systemically delineated research fields. The literature survey grasps the status of DH from a comprehensive and synthetic perspective. It is not only the technical advance in computational tools that is the main reason for the evolution of research paradigms towards frequency-driven models.

The predictive model proposed is presented here as a sample research task that – on the ground of the criteria proposed for delineation of the related terms and concepts – may be said to fit in the DH research paradigm. It was confirmed that binomials have a significant discriminative power for the classification of texts representing three categories of genres belonging to the company registration discourse. The findings show that the range of binomials (the structural types) that discriminate between various genres is wide, and these are not only the canonical sequences (the prototypical binomials, the true binomials, the proper binomials) that make up part of the predictive model. The conclusions related to the translational implications show evidence for the distortion of genre integrity in Polish translation of the texts. These findings are somewhat preliminary and they set clear vistas for further, more in-depth analysis in this respect, where the statistics on translation patterns would be complemented by findings from the predictive analysis conducted on the Polish parallel material.

References

Bhargava, Mudit/ Mehndiratta, Pulkit/ Asawa, Krishna (2013). "Stylometric analysis for authorship attribution on Twitter". In: Bhatnagar, V./ Srinivasa, S. (eds.) *Big Data Analytics. Second International Conference, BDA 2013 Mysore, India, December 2013 Proceedings*. Berlin. Pp. 37–48.

Biber, Douglas/ Finegan, Edward (eds.) (1994). *Sociolinguistic Perspectives in register*. New York.

Coyotl-Morales, Rosa María/ Villaseñor-Pineda, Luis/ Montes-y-Gómez, Manuel/ Rosso, Paolo (2006). "Authorship attribution using words sequences". In: Martínez-Trinidad, J. F./ Carrasco-Ochoa, J. A./ Kittler, J. (eds.) *Progress in Pattern Recognition, Image Analysis and Applications*. Berlin. Pp. 844–853.

Creswell, John W. (2013). *Projektowanie badań naukowych. Metody jakościowe, ilościowe i mieszane*. Kraków.

Dobrić Basaneže, Katia (2018). "Binomials in EU competition law". In: Marino, S./ Biel, Ł./ Bajčić, M./ Sosoni, V. (eds.) *Language and Law.The Role of Language and Translation in EU Competition Law*. Cham. Pp. 225–249.

Does, Jesse de/ Niestadt, Jan/ Depuydt, Katrien (2017). "Creating research environments with BlackLab". In: Odijk, J./ Hessen, A. van (eds.) *Clarin in the Low Countries*. London. Pp. 151–165.

Fabiszak, Małgorzata/ Konat, Barbara (2013). "Zastosowanie korpusów językowych w językoznawstwie kognitywnym". In: Stelmaszczyk, P. (ed.) *Metodologie językoznawstwa. Ewolucja języka. Ewolucja teorii językoznawczych*. Łódź. Pp. 131–142.

Fang, Chengyu Alex/ Cao, Jing (2015). *Text Genres and Registers: The Computation of Linguistic Features*. Berlin/Heidelberg.

Freake, Rachelle/ Gentil, Guillaume/ Sheyholislami, Jaffelr (2011). "A bilingual corpus-assisted study of the construction of naitionhood and belonging in Quebec". In: *Discourse and Society* 22(2). Pp. 21–47.

Frontini, Francesca/ Boukhaled, Mohamed, Amine/ Ganascia, Jean-Gabriel (2018). "Approaching French theatrical characters by syntactical analysis: a study with motifs and correspondence analysis". In: Legallois, D./ Charnois, T./ Larjavaara, M. (eds.) *The Grammar of Genres and Styles: From Discrete to Non-Discrete Units*. Berlin. Pp. 118–139.

Geeraerts, Dirk (1997). *Diachronic Prototype Semantics. A Contribution to Historical Lexicology*. New York.

Geeraerts, Dirk (1983). "Prototype theory and diachronic semantics: a case study". In: *Indogermanische Forschungen* 88. Pp. 1–32.

Graën, Johannes/ Volk, Martin. (2021). "Binomial adverbs in Romance and Germanic Languages – A corpus-based study". In: Lavid-López, J./ Maíz-Arévalo, C./ Zamorano-Mansilla, J. R. (eds.) *Corpora in Translation and Contrastive Research in the Digital Age: Recent advances and explorations*. Amsterdam. Pp. 326–342.

Gogora, Andrej (2020). "Ako a prečo pomenovať to, čo robíme? Problém slovenského prekladu termínu 'digital humanities'". In: *Slovenská Literatúra* 67(6). Pp. 598–613.

Grabowski, Łukasz (2015). "O frazeologii z perspektywy językoznawstwa korpusowego. Przegląd głównych nurtów badawczych ostatniego dwudziestolecia w Wielkiej Brytanii i USA". In: Bralewski, D. (ed.) *Problemy frazeologii europejskiej X*. Łask. Pp. 23–48.

Granger, Sylviane/ Paguot, Magali (2008). "Disentangling the phraseological web". In: Granger, S./ Meunier, F. (eds.) *Phraseology. An Interdisciplinary Perspective*. Amsterdam. Pp. 27–50.

Kubacki, Artur/ Więcławska, Edyta (2022). Modelling comprehensibility in legal translation from a computational and empirical perspective: How practice meets expectations. *Socjolingwistyka* 36. Pp. 229–252.

Kučera, Dalibor (2022). "Application of clarin linguistic tools in psychological research". In: Fišer, D./ Witt, A. (eds.) *Clarin: The Infrastructure for Language Resources*. Berlin/Boston. Pp. 747–780.

Langacker, Ronald Wayne (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford.

Legallois, Dominique/ Charnois, Thierry/ Larjavaara, Meri (2018). "The balance between quantitative and qualitative literary stylistics: how the method of »motifs« can help". In: Legallois, D./ Charnois, T./ Larjavaara, M. (eds.) *The Grammar of Genres and Styles: From Discrete to Non-Discrete Units*. Berlin. Pp. 165–193.

Levshina, Natalia (2015). *How to do Linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam/Philadelphia.

Maci, Stefania/ Sala, Michele (2022). "Corpus linguistics and translation tools for digital humanities: an introduction. In: Maci, S./ Sala, M. (eds.) *Research Methods and Applications*. London. Pp. 11–16.

McIntyre, Dan/ Walker, Brian (2019). *Corpus Stylistics*. Edinburgh.

Mellinger, Christopher D. (2017). *Quantitative Research Methods in Translation and Interpreting Studies*. London/New York.

Moretti, Franco (2013). *Distant Reading*. London.

Neumann, Stella (2014). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin.

Nirkhi, Smita/ Dharaskar, Rajiv Vasantrao (2013). "Comparative study of authorship identification techniques for cyber forensic analysis". In: *International Journal of Advanced Computer Science and Applications* 4(5). Pp. 32–35.

Nyhan, Julianne/ Flinn, Andrew (2016). *Computation and the Humanities. Towards an Oral History of Digital Humanities*. Berlin.

Oakes, Michael P. (2009). "Corpus linguistics and stylometry". In: Lüdeling, A./ Kytö, M. (eds.) *Corpus Linguistics. An International Handbook*. Berlin/New York. Pp. 1–21.

Pęzik, Piotr (2013). "Paradygmat dystrybcyjny w badaniach frazeologicznych. Powtarzalność, repreoducka i idiomatyzacja". In: Stelmaszczyk, P. (ed.) *Metodologie językoznawstwa. Ewolucja języka. Ewolucja teorii językoznawczych*. Łódź. Pp. 143–160.

Vaičenonienė, Jurgita/ Kren, Michal/ Lušicky, Vesna/ Vandeghinste, Vincent (2024). "Transnational Research Infrastructure: A Journey Through Clarin Knowledge Centres". In: *Digital Humanities in the Nordic and Baltic Countries Publications* 6. P. 1.

Więcławska, Edyta (2021). "Quantitative distribution of verbal structures with reference to the authorship factor in legal stylistics". In: *Logic, Grammar and Rhetoric* 66(79). Pp. 147–165.

Więcławska, Edyta (2022). "Predictive analysis for text classification: discrete units in company registration discourse". In: *Białostockie Studia Prawnicze* 27(4). Pp. 229–252.

Więcławska, Edyta (2023a). "Approaching legal multinomials from the sociolinguistic perspective – insights into authorship-based distinctions". In: *International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique* 36. Pp. 1699–1715.

Więcławska, Edyta (2023b). *Binomials in English/Polish Company Registration Discourse. The Study of Linguistic Profile and Translation Patterns*. Göttingen.

Więcławska, Edyta (2025). "Predictive analysis for text classification and its translational implications: Legal binominals in the generic perspective". In: Kubacki A./ Sulikowski, P. (eds.) *New Research Paradigms in Translation Studies Translation Landscapes – Internationale Schriften zur Übersetzungswissenschaft* 10. Götingen. Pp. 125–136.

Wissik, Tanja/ Wessels, Leon/ Fischer, Frank (2022). "The DH course registry: a piece of the puzzle in Clarin's technical and knowledge infrastructure". In: Fišer, D./ Witt, A. (eds.) *Clarin: The Infrastructure for Language Resources*. Berlin/Boston. Pp. 389–408.

Edyta Więcławska
Uniwersytet Rzeszowski
Instytut Anglistyki
Zakład Współczesnego i Historycznego Językoznawstwa Angielskiego i Porównawczego
Al. mjr. W. Kopisto 2 B
35-315 Rzeszów
Poland
ewieclawska@ur.edu.pl
ORCID: 0000-0003-0798-1940