

THE CORPUS-BASED ANALYSIS IN LEGAL DISCOURSE STUDIES

Науковий вісник Ужгородського університету. Серія: Філологія.

Випуск 2 (42)

УДК 81'33 : 81'42 : 340.113

DOI:10.24144/2663-6840/2019.2(42).96-99.

Томенчук М. Аналіз корпусу у вивченні юридичного дискурсу; кількість бібліографічних джерел – 19; мова англійська.

Abstract. The article deals with the methods normally associated with corpus linguistics and their application in legal discourse analysis. Some of them can be effectively used by researchers of legal language. Our analysis is based on the empirical data of legal documents, namely contracts. We have analyzed how processes such as collocation and concordance help to identify common features of representation in legal documents as well as direct analysts to representative texts in order to carry out qualitative analysis. The article outlines a possible framework for application and successful usage of corpus approaches, techniques and methods by legal discourse researchers.

We demonstrate that the variety of legal texts may include various and quite different characteristics of legal genres, such as different modes (speech, writing) and production circumstances in which legal genres are usually formed, different participants and relationships among them, or different communicative aims. Nevertheless, these are not the only differential peculiarities of legal texts, but most of them differ considerably in terms of their linguistic characteristics which can be defined if corpora and quantitative methods are used as tools for a corpus-based investigation of legal documents. Among the useful corpus linguistic techniques and tools one can use for legal discourse studies are the keywords, the concordance, the collocation display.

Both register and genre perspectives have been integrated into the corpus-based study of contracts. In the genre perspective a macrostructure, i.e. format has been outlined. From a register perspective, several distinctive lexico-grammatical features have been defined in the language of contracts. The latter are characterized by long sentences, impersonal constructions, conjoined phrases and words (usually nouns) resulting in an exceptionally dense use of technical vocabulary, multiple negation, the use of shall, etc.

Key words: collocation, concordance, discourse, genre, legal texts, legal documents, register.

Setting the problem. It is no exaggeration to say that the last two decades have seen a huge increase of interest in studying the relationships between law and language. Despite the recognition that legal language is indeed heterogeneous, most linguistically-oriented studies have so far used the term 'legal language' as a convenient label for generalized functional variety, or register, of the modern English. Unfortunately, such labeling often implies that it is either stationary or homogeneous ignoring a great degree of variability of legal language and its constant evolution. Research efforts have been essentially directed at identifying factors that make legal language distinctive relative to general, non-specialized language. However, it looks like there are virtually no studies that would provide an explicit description of linguistic variation *within* legal language or a description of variation between legal language and other specialized languages.

The analysis of the studies. The research in this area has been carried out in different perspectives that accounts for its complex character. The ways of investigation range from genre and discourse analysis [Bhatia 2004; Shuy 2001], to semiotics [Jackson 1994; Kevelson 1989], modality [Gotti 2001; Lauridsen 1992], and forensic linguistics [Coulthard and Johnson 2009; Gibbons 2003; Olsson 2004]. The increase of interest in the studying of law and language has encouraged some scholars to promote the appearance of a new interdisciplinary field of legal linguistics covering a range of different, although related, areas such as legal terminology and lexicography, legal translation and interpreting, analysis of legal discourse, courtroom discourse, etc. [Galdia 2014; Williams 2005].

In the construction of their analytical frameworks linguistic investigations have relied on the concepts of register and genre. However, there is some ambiguity in identifying such seemingly obvious and related terms and no consensus has been reached so far. In many studies, one concept is adopted and used exclusively while the others are neglected. For example, the term is exclusively used in some studies conducted by scientists like Bhatia, Swales [Bhatia 2004; Swales 1990]. In other studies, however, register is the preferred concept [Biber, Finegan 2001]. But in most cases the categorization is made based on external criteria relating to the speaker's purpose in communication and terms are, consequently, used to refer to the same varieties of texts, like novels, biographies, book reviews, newspaper articles, editorials, etc.

There are, however, research studies in Functional Linguistics which clearly differentiate between register and genre. Martin argues that register and genre are on different "semiotic planes" [Martin 1985]. Genre is viewed as a social process in which participants belonging to a certain culture use language in predictable sequential structures to fulfil certain communicative purposes. Genres have been also perceived as "conventional instances of organized text" [Couture 1986, p. 80]. Registers, on the other hand, have been referred to as the "expression plane" of genre [Martin 1985] and they tend to be associated with typical linguistic choices within different genres. The use of 'genre' and 'register' may signal different methodological approaches. The genre perspective usually focuses on issues related to discourse communities, ideology and power, while the register-oriented

study tends to deal with characteristic lexico-grammatical linguistic features.

The aim of the article is to outline a possible framework for application and successful usage of corpus approaches, techniques and methods in both register and genre perspectives studies by legal discourse researchers.

The account of the basic material. Despite the absence of a general consensus concerning the use of register and genre, research on legal language has been more consistent in their use. Thus, by and large, the term *register* has been used to refer to a general kind of language associated with a domain of use, i.e. law. On the other hand, the term *genre* has been understood as referring to more specialized varieties, such as brief, statute, contract, judgment, textbook or academic essay. In our study we share the perception of genre and register by Gozdz-Roszkowski [Gozdz-Roszkowski 2011, p.20], who differentiates two methodological perspectives and defines most important methodological characteristics of registers and genres. In the register perspective, the analysis is usually based on a sample of text excerpts representative of a particular variety and the focus is on lexical and grammatical features which are frequent and which are widely distributed across this variety. Such typical features are then examined according to their functions in the situational context of the variety. In contrast, the genre perspective focuses on language characteristics which may occur only once in a text and which are usually located at a specific place in the text. These linguistic characteristics can be specialized, formulaic expressions crucial to the construction of a particular genre. As a result, an analysis is based on complete texts. The language features are conventionally associated with the genre. They conform to the cultural expectations of how a particular genre should be constructed. The same texts can be analyzed from both register and genre perspectives.

Law is analyzed from the legal-linguistic perspective in order to understand it as a discursive practice. Diverse conceptual bases of law and the formation of argumentation by combining different legal signs account for diversity in legal discourse. Legal messages may have different logical and textual structures; they may be perceived as rules, principles, provisions, decisions and others. They can be characterized more precisely as legal speech acts of justification, description of facts, interpretation, argumentation, translation and so on, or written sources such as statutes, court precedents, judgments, orders, etc. But their common characteristic feature is the fact that they are expressed linguistically and they are used in meaningful textual forms.

In legal linguistics it is underlined that law as a discursive practice can be approached efficiently through the scrutiny of its language. Law should be primarily scrutinized as a linguistic phenomenon. This approach imposes itself because it refers the material part of law, to the elements of social reality that are construed as law and which can be read or heard as language. As a result, it shows law as a social phenomenon where power is exercised with linguistic means.

It constantly deals with one feature of law; it focuses upon its discursiveness [Galdia 2014, p.25].

What is usually referred to as “legal language” represents an extremely complex discourse embedded in a bewildering variety of legal writings. The extraordinary diversity of legal discourse was pointed out by Stanislaw Gozdz-Roszkowski: “Legal discourse spans a continuum from legislation enacted as different levels (e.g. state, federal), judicial decisions (judgments, decrees or orders), law reports, briefs, various contractual instruments, wills, power of attorney, etc., academic writing (e.g. journals, textbooks), through oral genres such as, for example, witness examination, jury summation, judge’s summing-up, etc. to various statements on law reproduced in the media and any fictional representation of the foregoing” [Gozdz-Roszkowski 2011, p.11].

The multitude of legal texts differ not only in situational characteristics of legal genres, such as modes, either speech or writing, and circumstances in which legal genres are produced, participants and their relations, or communicative purposes, but legal texts differ first and foremost in their linguistic characteristics. Important linguistic differences among legal texts have been found even in texts created in the same mode, i.e. written, and which deal with roughly the same topic [Gozdz-Roszkowski 2011]. The concept of discourse is defined as “a kind of thematically constrained text corpus from which the researcher has to extract an ‘actual’ corpus of analysis” [Spitzmüller 2011, p.76]. Although the utility of using corpus linguistics approaches in discourse analysis has already been demonstrated, the corpus design issue has not been clearly defined yet. But nowadays corpus is almost always synonymous with electronic corpus, i.e. a collection of texts which is stored on some kind of digital medium and used by linguists to retrieve linguistic items for research or by lexicographers for dictionary-making [Lindquist 2009, p. 3]. Corpora can be used as source of illustrative examples for discourse studies which are basically qualitative. A number of different terms have been created to describe these various approaches: ‘*corpus-driven*’ if you start with as few preconceived theoretical concepts as possible, ‘*corpus-based*’ if you use corpora and quantitative methods to investigate a problem which is formulated within a particular linguistic theory (this is the most common type), and ‘*corpus-aided*’ or ‘*corpus-supported*’ if you use corpora mainly to find illustrative examples.

The useful corpus linguistic techniques and tools used for discourse studies include the lemmatizer, keywords, the concordance, the collocation display. The lemmatizer makes it possible for a researcher to group all the inflexional forms of a word into the lemma (search word itself, a word or phrase that is interpreted), e.g. *take – take, takes, taking, took*. The advantage of using this that then it is possible to create a concordance for the lemma rather than having to create concordances for each verbal form.

The ‘keyword’ analysis has been popularized by Mike Scott, who created the corpus analytic tool Word-Smith Tools [Scott 2006]. By means of this program, it is possible to find out which words are special for

a certain text compared with some norm. This is done by calculating statistically which words are more frequent (positive keywords) and less frequent (negative keywords) than expected according to the norm. The method can be used to investigate typical traits of any text or group of texts or genre.

The concordance is a concordance sorting facility which searches for definite words in a text and sorts them into lines. Concordances are commonly generated from computer-based corpora to provide an exhaustive listing of the use of a word in its immediate textual contexts. Concordance software enables researchers to identify patterns that exist in authentic language that are not easily identifiable from a casual inspection of the printed text. Researchers can therefore study the contexts in which particular words occur.

The collocate display is used to automatically sort concordance lines according to their different collocation patterns. This tool displays the words adjacent to a search word within a collocation 'horizon' (or span) which can go up to 25 words. Entries can be sorted according to their collocates on both left and right. The tool is fast and very user-friendly and is therefore a good starting-point for identifying collocates in computer corpora.

On the one hand, the corpus linguistic methods and techniques offer the researcher a substantial high degree of objectivity; that is, they make it possible for the linguists to approach the texts under analysis free from any preconceived or existing notions concerning their linguistic or semantic/pragmatic content. On the other hand, corpus-based analysis means not only having a computer to objectively count and sort themes and linguistic patterns while using statistical algorithms onto textual data, but subjective researcher input is usually involved at almost every stage of the analysis. The researcher, informed by the quantitative aspects, has to decide what is to be analyzed. The quantitative analysis might be helpful in defining which corpus-based processes are to be applied to the data, and what the limits of statistical significance should be taken into account. The researcher is the one who has to make sense of the linguistic patterns obtained through corpus-based processes, usually with reference to one or more theoretical frameworks. All in all, the use of corpus linguistics techniques is becoming increasingly popular in discourse analysis nowadays.

REFERENCES

1. Bhatia, V.K. (2004). *Words of Written Discourse. A Genre-based View*. Continuum International Publishing Group.
2. Biber, D and E. Finegan (2001). Register variation and social dialect variation: The register axiom. In Eckert, P. and Rickford, J.R. (eds.) *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press. – P. 235–267.
3. Coulthard, M. and A. Johnson (2009). *An Introduction to Forensic Linguistics. Language in Evidence*. London: Routledge.
4. Couture, B. (1986). Effective ideation in written text: A functional approach to clarity and exigency. In Couture, B. (ed.) *Functional approaches to writing: research Perspectives*, Norwood, NJ: Ablex. P.69–91.
5. Galdia, Marcus. *Legal Discourses /Marcus Galdia.*– Peter Lang GmbH, 2014. 447 p.
6. Gibbons, J. (2003). *Forensic Linguistics. An Introduction to Language in the Justice System*. Blackwell Publishing.
7. Gotti, M. (2001). 'Semantic and Pragmatic values of *Shall* and *Will* in early modern English Statutes'. In: GattinadDossena (eds.) *Modality in Specialized Texts*. Bern: Peter Lang, 89–112.

Conclusions. In the genre perspective, the study of legal documents, namely contracts, compiled into a corpus can help the researcher to define and analyze the expected textual conventions for complete texts of this type. The analysis of the convention of the texts of this type enables to specify the most important constituents of the contract. This type of analysis results in proposing a macrostructure, i.e. format outline. The contract should begin with the parties who enter into agreement, and then comes the recital in which the subject matter and main points of the contract are stated with some background information and definitions and interpretations of certain words and notions used within it. The main part of the contract includes the operative provisions, which may be categorized as warranties and conditions. There is usually the force majeure which states that the contact can be terminated in the event of some unexpected events outside of the control of the parties. A typical contract closes with the signatures section. Finally, there are some attachments, which may include more details to the contract, e.g. price list. All these conventional parts arranged in accordance with a specific organizational format contribute to the creation of what a legal culture recognizes as the genre of contract.

From a register perspective, several distinctive lexico-grammatical features have been defined in the language of contracts, such as, for example, the excessive use of the passive voice, conditionals, archaic adverbs and prepositional phrases, etc. Contracts are characterized by long sentences, impersonal constructions, conjoined phrases and words (usually nouns) resulting in an exceptionally dense use of technical vocabulary (e.g. *claim, loss, damage, liability, remedy or action*) multiple negation, the use of *shall*, etc. The contractual provision is marked by relatively few verb phrases and a heavy reliance on phrasal syntax. Cohesion in the contract is ensured through repetition of lexical items (e.g. *Client, Holder, Lawyer*). The frequent Phrasal coordination and past participle forms placed in post-nominal position (e.g., *registration and qualification effected pursuant to...*) are frequent. The frequent use of the determiner *any* accounts for all conceivable contingencies in contracts.

Thus both register and genre perspectives can be integrated into the corpus-based study of legal documents.

8. Gozd-Roszkowski, Stanislaw. Patterns of Linguistic Variation in American Legal English: A Corpus-based Study. EBSCO Publishing. Peter Lang AG. 2011. 4
9. Jackson, B. (1994). Legal Semiotics and the Sociology of Law. Onati International Institute for Sociology of Law.
10. Kevelson, R. (1989). The Law and Semiotics. Dordrecht: Kluwer/Plenum Press.
11. Lauridsen, K.M. (1992). 'The meaning and Use of the Modals *Can* and *May* in English Contract law texts. *Hermes, Journal of Linguistics* 9, 43–64.
12. Lindquist Hans (2009). Corpus Linguistics and the Description of English. Edinburgh University Press. 220 p.
13. Martin, J.R. (1985). Process and Text: Two aspects of human semiotics. In: Benson, J.D. and Greaves, W.S. (eds.) Systemic Perspectives on Discourse (Vol.1). Norwood, NJ: Ablex. P.248–274.
14. Olsson, J. (2004) Forensic Linguistics: An Introduction to Language, Crime and the Law. New York: The Continuum.
15. Scott, Mike (2004). WordSmith tools, Version 4.0. Oxford University Press.
16. Shuy, R.W. (2001). 'Discourse Analysis in the Legal Context'. In: Schiffrin, D., Tannen, D. and Hamilton, H.E. (eds). The handbook of Discourse Analysis, Blackwell Publishing, 437–452.
17. Spitzmüller, Jürgen and Warnke, Ingo H. (2011). Discourse as a 'linguistic object': methodical and methodological delimitations. *Critical Discourse Studies*, 8: 2. P.75–94.
18. Swales, J. (1990). Genre Analysis. Cambridge: Cambridge University Press.
19. Williams, C. (2005). Tradition and Change in Legal English. Frankfurt am Main: Peter Lang.

АНАЛІЗ КОРПУСУ У ВИВЧЕННІ ЮРИДИЧНОГО ДИСКУРСУ

Анотація. У статті розглядаються методи, які зазвичай асоціюються з корпусною лінгвістикою, та можливість використання їх для вивчення юридичного дискурсу. Деякі з методів дослідники ефективно застосовують під час аналізу юридичної мови. Наше дослідження ґрунтується на текстах юридичних документів, зокрема контрактів. Ми з'ясували, що такі процеси, як сполучення слів та узгодженість між ними, допомагають визначити спільні риси, які характеризують мову юридичних документів, і водночас спрямовують дослідників на здійснення кількісного аналізу. Стаття окреслює для дослідників юридичного дискурсу можливу структурну модель застосування та успішного використання деяких підходів, технік та методів корпусної лінгвістики.

Різні типи юридичних текстів можуть мати цілком відмінні характеристики багатьох юридичних жанрів, зокрема різні форми (усну, письмову), обставини, за яких вони були сформовані, різних учасників і відносини між ними та різні комунікативні цілі. Незважаючи на це, не тільки такі характеристики становлять визначальні відмінності юридичних текстів, а також більшість із них значно відрізняється щодо лінгвістичних особливостей, які можна визначити, застосовуючи методи та техніки корпусної лінгвістики. Серед корисних і продуктивних технік та інструментів для вивчення саме юридичного дискурсу можна відзначити ключові слова, сполучення слів та ряди узгодженості між словами.

Обидва плани реєстру та жанру було залучено до корпусного аналізу контрактів. Із жанрової перспективи виділено макроструктуру контракту, тобто його формат. У плані реєстру було визначено основні лексико-граматичні риси контрактів. Останні характеризуються довгими і складними реченнями, безособовими конструкціями, об'єднаними фразами і словами (зазвичай іменниками), використання згаданих засобів спричиняє щільне уживання технічного вокабуляру, подвійні заперечення, використання *shall* тощо.

Ключові слова: сполучення слів, узгодженість, дискурс, жанр, юридичні тексти, юридичні документи, реєстр.

Стаття надійшла до редакції 30 вересня 2019 р.

© Томенчук М. 2019 р.

Мар'яна Томенчук – кандидат філологічних наук, доцент кафедри прикладної лінгвістики Ужгородського національного університету, Ужгород, Україна; <https://orcid.org/0000-0002-2036-4616>

Maryana Tomenchuk – Candidate of Philology, Associate Professor of the Applied Linguistics Department, Uzhhorod National University, Uzhhorod, Ukraine; <https://orcid.org/0000-0002-2036-4616>