DISCUSSION PUBLISHERS
Research Center · Open Access Edition

Corpus Linguistics in Legal Discourse

Stanisaw GodRoszkowski

Discussion Research ID	DPR-2026-1332
Year	2021
Field	Social Sciences
DOI	10.1007/s11196-021-09860-8
Author ORCID	0000-0002-4323-8647
Licence	CC-BY
Original source	International Journal for the Semiotics of Law - Revue internationale de Smiotique juridiq
Permanent link	https://discussionpublishers.com/research/319e6260-9e30-4304-87f1-34d2d53cf890
Publisher	Discussion Publishers discussionpublishers.com

Abstract

Abstract There are many different ways in which modern Corpus Linguistics can be used to enrich and broaden our understanding of legal discourse. Based on the central principle of co-occurrence and co-selection in language construction, this paper reviews current applications of Corpus Linguistics in the area of legal discourse focusing on issues ranging from phraseology, variation in legal discourse, legal translation, register and genre perspectives on legal discourse, legal discourse in forensic contexts to evaluative language in judicial settings. It revisits the notion of corpus and it highlights the relevance of various types of legal corpora and computer tools in legal linguistic research.

Republished by Discussion Publishers under the original CC-BY licence. Copyright remains with the original authors; please cite the original work.

Cite as: Stanisaw GodRoszkowski (2021). Corpus Linguistics in Legal Discourse. Discussion Publishers, DPR-2026-1332.

Int J Semiot Law (2021) 34:1515–1540
https://doi.org/10.1007/s11196-021-09860-8

Corpus Linguistics in Legal Discourse

Stanisław Goźdź-Roszkowski[1]

Accepted: 19 July 2021 / Published online: 2 September 2021
© The Author(s) 2021

Abstract
There are many different ways in which modern Corpus Linguistics can be used to enrich and broaden our understanding of legal discourse. Based on the central principle of co-occurrence and co-selection in language construction, this paper reviews current applications of Corpus Linguistics in the area of legal discourse focusing on issues ranging from phraseology, variation in legal discourse, legal translation, register and genre perspectives on legal discourse, legal discourse in forensic contexts to evaluative language in judicial settings. It revisits the notion of 'corpus' and it highlights the relevance of various types of legal corpora and computer tools in legal linguistic research.

Keywords Corpus linguistics · Legal discourse · Phraseology · Variation · Evaluation · Forensic linguistics

1 Introduction

Corpus Linguistics has revolutionized the way language is understood and explored today leading to a proliferation of empirical studies on virtually any aspect of language. The explosion and expansion of corpora and computer tools in the fields of descriptive and applied linguistics have been made possible mainly because of the information revolution of the late twentieth century [76]. What started as a methodological enhancement secured by increasing quantities of data, enriched and processed by ever more powerful and efficient computers, has emerged as a theoretical and qualitative revolution which "has offered insights into the language that have shaken the underlying assumptions behind many well-established theoretical positions in the field" [104: p.17].

It soon became very clear that Corpus Linguistics has had much to offer many other areas apart from linguistics, especially those where language and other

✉ Stanisław Goźdź-Roszkowski
stanislaw.gozdz@uni.lodz.pl

[1] Department of Specialized Languages and Intercultural Communication, University of Łódź, Pomorska 171/173, 90-236, Lodz, Poland

Discussion Publishers edition DPR-2026-1332

disciplines are intimately bound up. This is particularly true of the disciplinary discourse of law, where language is central to its construction and interpretation. Interestingly, recent years have seen an unprecedented growth of corpus-informed research into legal discourse, with its traditional interest in areas such as variation, phraseology, translation, terminology or phraseology, only to be paralleled by explorations carried out by both legal academics and practitioners embracing corpus linguistics methods as a new tool [91, 93, 107]. These are mainly concerned with legal interpretation, especially regarding the ordinary meaning of terms.

While Corpus Linguistics may offer an exciting prospect for both lawyers and linguists of a fruitful relationship with real-life and big data to uncover patterns of language use in law and legal texts which would otherwise go undetected, there are also pitfalls and limitations one should be aware of. Accordingly, this paper will explore some issues around the use of Corpus Linguistics in analyzing legal discourse, including both its advantages and also some of the methodological challenges associated with its use [6]. It appears that there are certain aspects of legal discourse which are particularly amenable to corpus analysis, such as its inherently formulaic nature, and numerous corpus-based studies have explored this issue, especially in the context of variation. The methodology of Corpus Linguistics is inherently quantitative (statistical) but frequency of lexical items may become revealing only if a comparative perspective is adopted. This means that linguistic analyses tend to characterize legal discourse, or its certain aspects, by comparing it with other specialized or general discourses or by comparing different types of legal discourse with one another.

Given the sheer volume of current corpus-informed work being done in a wide range of legal languages and cultures, this overview needs to be necessarily restrictive in both depth and scope. It will prioritize phraseology in its current broad sense, as a complex and multifarious phenomenon central to language description and its organization [59]. Thanks to the commonly accepted usage-based, context-of-use and inductive approaches found in contemporary Corpus Linguistics, we know now that much more language is locked in 'fixed phrases' than previously thought [90]. This is particularly true of legal discourse in its various textual manifestations. The research done in corpus linguistics phraseology has had a profound impact on other (sub)areas, such as formulaicity and standardization in legal documents, legal terminology, variation within legal discourse, legal translation, authorship attribution, to name only few. In this sense, corpus linguistics research in these sub-areas is a derivative of research into corpus linguistics phraseology and it will be a key focus of attention throughout this paper.

Finally, it needs to be pointed out that not all use of computerized data and computer tools should be associated with Corpus Linguistics. Computational approaches to argumentation in law are just one obvious example. Corpus investigation tools are being increasingly used by researchers who would not describe themselves as corpus linguists or even as linguists. The underlying premise of this paper is that the inclusion of the designation 'corpus linguistics' presupposes a central concern with how the linguistic and legal perspectives can be interconnected to shed more light on legal discourse, while keeping a clear focus on the language perspective of the law and language interplay. Recently, a new field of study called Computer-Assisted

Discussion Publishers edition DPR-2026-1332

Legal Linguistics has been proposed [107] in an attempt to reconcile the different linguistic and legal applications of using corpora and computer tools. However, this article is in no way confined to research subsumed carried out under the mantle of either 'Legal Linguistics' or 'Computer-Assisted Legal Linguistics.

2 What is Corpus Linguistics Today?

Corpus Linguistics has become a term referring to a wide range of activities and approaches [e.g. 60]. What appears common to all of them is collecting large quantities of text in electronic form so that they are open to data-manipulation techniques. Corpus Linguistics could then be succinctly defined as "[t]he set of studies into the form and/or function of language which incorporate the use of computerized corpora in their analysis" [82, p. 5]. This broad definition reflects an enormous body of research carried out using techniques focusing on a search term and observing its immediate environments (key-word-in-context or concordance lines); calculating relative frequency (e.g. collocation studies) annotating such categories as word class, grammatical function or semantic class to carry out frequency calculations based on such categories. Corpora have been explored to analyze different units of language, ranging from single lexical items (studied, for example, in lexicology and dictionary meaning), collocations and typical phraseologies of a discourse type, through the level of clause and lexicogrammar to the level of an entire text. This broadening of scope in terms of a unit of analysis has led to extensive explorations in textual cohesion, authorial style, figurative meaning, evaluative meaning, legal, social, political, cultural and religious ideologies encoded in text, and much else.

In addition, contemporary Corpus Linguistics has become highly diversified transforming into various forms depending on the researcher's approach towards the status and use of corpus data. How corpus evidence is treated by researchers has in fact become a major issue resulting in at least three different approaches [cf.105: pp. 84–87]:

- **Corpus-based approach** – this approach aims to verify existing claims arrived at through non-Corpus-Linguistics methods (for example, intuition, serendipity, qualitative analysis, such as a close reading of a text). This type of analysis starts with a hypothesis which is then proved or refuted based on corpus evidence. This has been traditionally one of the most common approaches.
- **Corpus-driven approach** – according to this approach, corpus is treated as the sole source of information to explore a phenomenon with no prior assumptions made. This inductive method of analysis has led to numerous studies of legal phraseologies by means of the so-called 'lexical bundles'. It is worth consulting [66] for a representative example of a corpus-driven diachronic approach adopted to investigate standardization in legal and administrative texts.
- **Corpus-assisted approach –** Partington et al. [82: p.10] define Corpus-assisted Discourse Studies (CADS) as a subset of Corpus Linguistics in the following manner: „[s]et of studies into the form and/or function of language as *communicative discourse* which incorporate the use of computerised corpora in their anal-

Discussion Publishers edition DPR-2026-1332

yses". This approach is eclectic and the researcher is encouraged to draw upon various analytical techniques, apart from the corpus methodology, in an effort to achieve the desirable results. It seems that this approach can be noticed, especially among legal scholars who tend to use corpus material to explore issues formulated from the legal dogmatic perspective [e.g. 71].

As will become clear in the following sections, each of these approaches, albeit to a varying extent, has been adopted to investigate various aspects of legal discourse. The corpus-assisted approach marks an emerging characteristic of current Corpus Linguistics, which is that it supports and is supported by other approaches to language. For example, Potts and Kjaer [87] combine Corpus Linguistics methods with Critical Discourse Analysis [30] to explore how 'achievements' are discursively constructed in the texts created by the International Criminal Tribunal for Yugoslavia. The combination includes quantitative corpus linguistic methods, which help to quantify, visualize and detect patterns of meaning in data, and a qualitative theory of Critical Discourse Analysis useful for shaping and informing the explorations of the quantitative results. As Potts and Kjaer [87: p. 530] aptly comment, most corpus-informed work is largely for descriptive purposes and it may lack the critical view on language use in the legal field.

Thus, Corpus Linguistics is becoming a versatile methodology for language analysis in legal contexts, which welcomes mixing and triangulating various approaches to achieve the most relevant and complete results. One consequence of this trend is that the problem for the researcher has shifted from accessing large enough quantities of data (as these have become relatively easy to obtain thanks to publicly available resources) to elaborating a range of reliable and consistent analytical methods informed by an appropriate theoretical framework, all geared towards addressing specific, often interdisciplinary (legal linguistic), research issues.

2.1 Legal Corpora and Tools for Analyzing Corpora

2.1.1 The Corpus

A corpus is simply a collection of texts stored according to specific criteria, which can be processed using specialist software [114]. The criteria routinely involve issues such as sampling and representativeness, finite size, machine-readable form, and authenticity [77: pp. 4–5]. Recently, the notion of a corpus seems to have been expanded to include any digitally stored collection of texts. This can be noticed, especially in the legal applications of corpora, where less attention is given to the linguistic criteria. It is worth pointing out that it was linguists who have had a central role in developing corpora and shaping the technology in ways best suited to their needs. This point should be borne in mind when considering the use of existing corpora and computer tools for interdisciplinary research involving legal scientists and their research agenda. As Solan & Gales [93: p.1312] rightly point out "[thus] corpora and the tools for using them are most likely to assist in addressing legal issues when the law considers the distribution of language usage to be

Discussion Publishers edition DPR-2026-1332

legally relevant". But this raises the question of whether criteria used by linguists are equally relevant for lawyers? Can the tools be used for interdisciplinary research agenda? While this paper addresses primarily linguistic concerns, answers to these questions can be explored by reflecting upon the purpose for creating a corpus and its design criteria. In fact, it seems that most corpora are created for the purpose of a specific project [107]. For example, many corpora have been created to facilitate the process of translating legal texts, as well as to provide data for researching legal translation (see [86] for a recent overview of corpora and research methods in legal translation studies).

The reason for compiling a corpus may impact the level of corpus annotation, i.e. adding information to the texts collected. There are two main types of annotation: part-of-speech (POS) tagging and mark-up. In the former, each lexical element in the corpus is assigned a label or a tag indicating its grammatical status, i.e. noun, verb, adverb, etc. The latter involves indicating structural units, such as introductions or closing sections in a document. In a transcript of a trial, information could be added about speech turns, pauses or paralinguistic features signaling the use of body language. The POS tagging is now routinely carried out (semi) automatically using a specialist software but mark-up requires manual work and it can be extremely time-consuming.

2.1.2 Corpora in Legal Discourse

Legal corpora are predominantly text corpora but collections of video data annotated with information about, for example, gestural expressivity, dialogue functions, etc. could be very helpful in facilitating pragmatics research into, for example, courtroom discourse. Unfortunately, such multi-modal corpora in legal discourse are very scare, especially in comparison with other domain-specific or general languages. The interested reader is advised to browse the website of the CLARIN project featuring cutting-edge research infrastructure for language as social and cultural data, including multi-modal data [20].

As is well known, legal discourse spans a continuum from legislation enacted at 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 above-mentioned [42]. The extraordinary diversity of legal discourse is, to some extent, reflected in the growing number of legal corpora and their different types.

However, the availability of legal resources may vary dramatically given the inherently confidential and private nature of some legal documents and the institutional confines within which they are created. There are several useful overviews of contemporary legal corpora [72, 85, 86, 107]. Worth recommending is also SOULL (Sources of Language and Law), an open on-line platform, regularly updated to provide a wealth of information about existing data collections and copora of legal language [94].

Discussion Publishers edition DPR-2026-1332

Many publicly available legal corpora have been created under the auspices of international institutions such as EU, UN or WTO. Examples include the *United Nations Parallel Corpus* [116], the *JRC-Acquis Parallel Corpus* [95], the *Digital Corpus of the European Parliament DCEP* [57]. The names of the corpora are evident of a bias towards multilingual and translated legislation. However, the range of legal corpora is rapidly increasing, with both new legal genres and legal languages being constantly added. For example, there is a recent project run by dr Karen McAuliffe with the goal of creating language resources for investigating judicial discourse at the European Court of Justice [69].

It should be pointed out that relatively small, carefully-targeted corpora (ranging from less than a milion words up to a few milion words) could be extremely effective in describing particular types of legal discourse. These are often compiled by researchers for the purpose of a specific research task. For example, Wei Yu [109] compares patterns of distribution and use of reporting verbs in self-compiled corpus of UK Supreme Court court judgments (of 1.17 million) with the BNC sampler corpus as a reference corpus (1.01 million) using an existing programme *Wmatrix* [88]. In this study, reporting verbs are treated as 'entry points' providing insights into the construction of argumentation in judicial opinions.

Such datasets may not always adhere to a rigorous sampling frame. Instead, they may contain whatever language data is available (see, for example Zaśko-Zielińska [115] for a description of the *Polish Corpus of Suicide Notes* (PCSN). Self-made corpora can be then compared with pre-existing large corpora, whether *heterogeneric* corpora, such as the *Corpus of Contemporary American English* or *monogeneric* corpora (e.g. Corpus of US Supreme Court Opinions).

Other types of corpora have been usually described in terms of dichotomies: *general* vs. *specialized*, *monolingual* vs. *bi-or-multi-lingual*, *comparable* vs. *parallel*, *diachronic* vs. *synchronic*, etc. (see [85] for an overview of legal corpora in multilingual settings). The choice of a specific type of corpus depends on research goals but discourse analysis is inevitably comparative. As already indicated, it is only possible to describe a type of legal discourse by comparing it with other discourse types, legal or non-legal. For example, Whittaker [110] uses bilingual corpora of French and Norwegian texts to illustrate how two different legal cultures may vary, especially in terms of the degree of terminologisation as well as stylistic features of legal discourse. Leung [70] draws upon two sets of bilingual data of appproximately 800,000 words representing two different genres of legislation (the Sexual Offences related ordinances of the Criminal Act) and transcripts of the proceedings of cross-examinations by the counsel in five separate rape trials recorded on about 101 audiotapes in order to scrutinise practices of the bilingual, legal system and the impact of bilingual legislation, translation, and interpretation on trial proceedings.

2.1.3 Tools for Analyzing Corpora

It is often pointed out that no matter how good a corpus can be in terms of representativeness, balance, size, etc., it is practically worthless without a suitable software to explore it. Even if software development has been constantly improved, its central functions have remained relatively stable. The most widely used tools for

Discussion Publishers edition DPR-2026-1332

interrogating corpora are still the concordancer and various "calculators of *frequency, keyness, clusters* and *dispersion* [82: p. 17]. They are all usually found in software packages such as, for example, *WordSmith Tools* developed by Mike Scott or *AntConc* described as a freeware corpus analysis toolkit for concordancing and text analysis. Worth recommending is also *Sketch Engine*, which also provides access to a range of multi-lingual resources (most of which are non-legal) and provides tools for looking up translations and creating bilingual glossaries of terminology.

Such tools are indispensable if the researcher intends to create her own unique dataset. In case of existing legal corpora, there are usually some in-built search options. For example, the *Corpus of US Supreme Court Opinions* [22] offers a range of features enabling one to see the frequency of words and phrases, identifying and comparing collocates, examining words and phrases in their contexts and comparing them in different periods. This shows that the patterns of repetition and patterns of co-selection have become a significant element in a corpus allowing for detecting non-obvious patterns.

The following sections will demonstrate how standard tools of querying corpora have been used to add to our understanding of legal corpus.

3 Linguistic Applications

This overview starts by considering how corpus linguistics methodology has been applied in legal discourse adopting traditional and well-established linguistic perspectives, such as phraseology, variation, register and genre studies, forensic linguistics and translation. It will then move on to address the lesser known area of evaluative language in legal, especially judicial discourse.

3.1 Phraseology in Legal Discourse

3.1.1 The Notion of Phraseology in Corpus Linguistics

Corpus linguistics research has revolutionized the way phraseology is now understood and studied in general language as well as in its specialized varieties. Some of the ground covered in this section can be found in earlier overviews [41, 43].

It is instructive to compare two different definitions, which are indicative of two radically different approaches to phraseology. In the earlier, non-corpus one, phraseology is defined as "[the] structure, meaning and use of word combinations" [27: p. 3168]. In the more recent one, Hunston [60, p. 5] refers to it as"a very general term used to describe the tendency of words, and groups of words, to occur more frequently in some environments than in others".

The former represents a traditional approach which views phraseology as a continuum along which word combinations are situated, from the most opaque and fixed ones to the most transparent and varied. This research tradition aims at identifying linguistic criteria for distinguishing one type of phraseological unit

Discussion Publishers edition DPR-2026-1332

from another, with the most idiomatic units (idioms, proverbs) presented as the most 'core'. It is worth consulting Granger and Paquot [46] and, more recently, Pęzik [84] for a useful overview of the traditional (analytical) and corpus-linguistics (distributional) approaches, their major characteristics and limitations. The classical paradigm has generated sophisticated analytical typologies of phraseological units.

The latter marks a new research trend with an emphasis on"flexible word sequences which display some consistency in aspects of form but more so in aspects of meaning" [60: p. 5]. Consequently, 'corpus linguistics phraseology' (a term now used interchangeably with 'distributional phraseology' or 'frequency-driven phraseology') prioritizes differential frequencies as a way to identify patterns of repetition and patterns of co-selection as significant elements in a corpus of texts under study. This means that studies of phraseology should not only pay attention to lexical and grammatical co-occurrence (collocation and colligation, respectively) but also to textual co-occurrence, when lexical items occur differentially in different parts of a text, such as in paragraph or text initial position. Questions as to whether such differential frequencies are register or genre independent, or, conversely, genre or register-specific have also informed research into phraseology in legal texts.

There are six criteria useful in defining the object of phraseological research [47: p. 4]. First, the nature of the elements involved in a phraseologism should be considered to decide whether the analytical focus should be on wordforms, lemmas and grammatical forms (or both) or perhaps on broadly-defined elements of meaning, such as semantic sequences (see Sect. 3.1.4 for an example of investigating semantic sequences in judicial discourse). Second, the number of elements involved in a phraseological units could range from two elements (as in typical studies of collocation) to eight (as sometimes found in studies of lexical bundles (see Sect. 3.1.2). Third, in contrast to early studies of phraseology, where frequency was not of paramount importance, corpus studies prioritize items which must be observed more frequently than expected. Fourth, explorations of phraseology in legal discourse focus on contiguous elements allowing no distance (e,g. lexical bundles, n-grams, clusters) or on discontinuous elements (e.g. skipgrams, phrase frames). Fifth, the degree of lexical and syntactic flexibility of the elements involved could also be considered. Finally, the sixth criterion concerns semantic unity and semantic non-compositionality of phraseologisms. Research on legal phraseology has largely focused on very frequent and recurrent but semantically transparent items. The six criteria provided above help to broaden the scope of phraseological research by extending the range of linguistic items that could be classified as phraseologisms. At the same time, they ensure greater rigour in analyzing word combinations Gries [47: p. 6] provides the following, oft-cited definition of phraseologism:

> The co-occurrence of a form or a lemma of a lexical item and one more or additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance.

In the next section, we turn to consider some of the common methods of exploring phraseology in legal discourse.

Discussion Publishers edition DPR-2026-1332

3.1.2 Approaches to the Study of Phraseology in Legal Discourse

It should be borne in mind that research into legal phraseology has been influenced by both the classical and analytical approach, as well as the more recent corpus perspective. Regarding the corpus studies, they reflect the general distinction between corpus-based and corpus-driven methodologies [105: pp. 84–87]. The corpus-based approach involves pre-selecting specific expressions and then searching for their frequencies, instances of use, and enabling comparisons across different text genres in case of a multi-genre dataset. For example, Koźbiał [67: pp. 495–542] in his comprehensive study of EU and Polish judicial discourse, sets out to find out if, and how, Polish judges employ Latin expressions in their discourse. This was accomplished by first creating a pre-defined list of Latin terms, phrases and maxims based on external, non-corpus sources. The final list was then fed into the *WordSmith Tools* software to generate lists of concordances which were saved as Excel spreadsheets to automatically segregate the data. The quantitative data was subsequently used to study distribution of Latinisms in the corpora of CJEU and national judgments. The corpus-based approach can be used to check the relevance of findings found in early, non-corpus studies, as well as to verify intuitions and hypotheses about appropriacy and typicality of particular expressions in a given legal genre.

In contrast, the corpus-driven approach is both inductive and inclusive by working bottom-up to provide a wide range of word sequences. Using a number of bespoke computer programmes (e.g. *kfNgram*, *ConcGram*, *WordSmith Tools*, etc.) it has now become possible to generate various contiguous (e.g. lexical bundles, n-grams) or non-contiguous (e.g. skipgrams [19]; phrase frames [35]) recurrent word combinations which elude the traditional predefined linguistic categories characteristic of the corpus-based approach. These linguistic constructs vary in terms of the permissible distance between two associated words and the number of variants allowed. For example, concgrams represent a much more flexible approach to the co-occurrence of two or more words because they represent co-occurrences which take into account "all the configurations of the co-occurring words irrespective of any constituency and/or positional variation" [108: p. 116]. For example, a concgram created on the basis of *increase* and *expenditure* could be instantiated as *increase in expenditure, increase in the share of expenditure* (constituency variation) and *expenditure would inevitably increase* (positional variation). It turns out that contractual discourse may be particularly amenable to this type of analysis (see, for example, Basaneže [4] for a study of concgrams in various contractual instruments).

It is the lexical bundles that have turned out to be the most researched type of multi-word units with numerous applications in such areas as variation, standardisation, translation, discourse function in judicial argumentation, etc. Identified on the basis of frequency alone, these are uninterrupted word sequences of varying length (usually studied as sequences of four elements) which are semantically transparent. Even if not perceptually salient, they are considered as important building blocks in discourse providing a kind of discourse frame for expressing new information [10]. They have been usually analysed in terms of their structural (depending on the dominant part of speech, e.g. noun, verb, a clause fragment) and functional properties. Biel [12: p. 11] provides the following examples of lexical bundles in EU law:

Discussion Publishers edition DPR-2026-1332

referred to in Article, in accordance with the, of regulation EU No., having regard to the, for the purposes of, the European Parliament and, Member States shall ensure that. However, one major limitation of lexical bundles is that they are inflexible in terms of the permissible distance between its constituent elements. It is then advisable to rely on other types of multi-word expressions to obtain a full phraseological picture of a given legal genre or genres.

The influence of corpus linguistics methodology on how legal phraseology has been investigated extends beyond technological advances in text processing. Rather, corpus linguistics phraseology has paved the way for new and innovative studies which have begun to reveal the potential for investigating various roles and functions performed by different multi-word units in legal discourse.

3.1.3 The Applications of Corpus Approaches to Phraseology in Legal Discourse

There have been a number of studies which use the corpus-driven approach to construct classifications of recurrent word combinations in accordance with their functional properties. Goźdź-Roszkowski's cross-genre classification of legal lexical bundles [42: pp. 109–142][1] has revealed that legal genres can be described and differentiated in terms of their preferred phraseologies. These phraseological preferences correlate strongly with the different communicative priorities and epistemological precepts of the legal genres. Legal genres seem to be characterized by a varying degree of formulaicity (the occurrence of fixed and recurrent expressions). While all the genres under study use lexical bundles, they use them to differing extents, and for different functions. Operative genres (legislation and contracts) are structured around lexical bundles to a greater extent than any other genre. In sum, this typology was used to add to our understanding of variation in legal discourse.

In a similar vein, Biel's classification of patterns based on a phraseological continuum is used to shed light on variation in the primary genre of legislation [13: pp. 178–182],[2] while Kopaczyk's taxonomy of short and long bundles in the early legal discourse of Scottish burghs [66] is a fine example of a diachronic study using phraseological chunks of language to trace standardization patterns in Scots legal and administrative texts.

Given the central importance of phraseology and terminology for legal translation (e.g. [11, 89]), it is not surprising that lexical bundles have been also used to investigate the impact of translation process on the patterning of legal language. A study reported in Biel [12] explores internal variation in legislation relying on different legal corpora: a bilingual corpus of translator-mediated EU legislation (regulations and directives) – the Polish Eurolect corpus, compared against two reference corpora: the English Eurolect corpus and the corpus of non-translated Polish legislation

[1] (1) *Legal reference*: temporal, location, attributive, participative, institutional, terminological, procedure-related bundles; (2) *text-oriented*: causative/resultative, condition, clarification/topic elaboration, focus, framing, structuring, transition bundles; (3) *stance*: epistemic and attitudinal bundles.

[2] (1) Text-organising patterns; (2) grammatical patterns; (3) term-forming patterns (multi-word terms); (4) term-embedded collocations; (5) lexical collocations.

Discussion Publishers edition DPR-2026-1332

(Polish Domestic Law Corpus). This is an excellent example of how different types of corpora (multilingual vs monolingual; translator-mediated vs. comparable) can be utilized to "account for two fundamental relations of translations: the relation to source texts and the relation to non-translated target-language texts of a comparable genre" [12: p. 15]. A similar approach combining translation and comparative perspectives is adopted in Koźbiał [67] who uses lexical bundles to investigate the formulaic profile of EU and national judgments. A functional classification of lexical bundles in EU and Polish judicial discourse is presented to assess the degree of (dis) similarity between translated EU judgments and non-translated national judgments.

The description of regularity in judicial texts from the perspective of Argumentation Theory [33] is also the object of a study reported in Mazzi [73] which draws on 4- and 5-gram frequency lists to search a corpus of Supreme Court of Ireland's judgments as a first stage in an investigation of the discourse and use of causal argumentation around the theme of data protection. This study is instructive in that it skilfully combines a quantitative analysis of recurrent phraseology with a qualitative study of the judgments where the usage patterns established earlier were observed to be most frequent. This led to detecting and reconstructing textual sequences embedding causal argumentation.

These studies have confirmed the relevance of corpus linguistics in discovering recurrent phraseological patterns, especially those that elude the earlier traditional classifications of legal phrasemes [64]. There are several different research trends exploiting corpus methods which could be roughly grouped into (1) research into collocations; (2) terminologically or terminographically-oriented research; (3) research into routine formulae (including binomials or trinomials); (4) cross-language studies. Future research should integrate, to a larger extent, corpus analyses of various multi-word units with explorations of their institutional contexts, as well as their social and cognitive foundations. One example of such studies is [79] which explores how binomials and multinomials can structure our social experience through their use and reproduction in legal documents.

3.1.4 Extending the Concept of Phraseology: Semantic Sequences in Judicial Discourse

The concept of 'semantic sequence' is a relatively recent and still very promising candidate for a novel way of describing textual recurrence in legal discourse. Its origin lies with a corpus linguistics research carried out in a series of works by Hunston (e.g. [60]). It reflects a much broader understanding of what could be the object of phraseological research because semantic sequences refer to meaning elements rather than only co-occurrences of wordforms or lemmas.

Goźdź-Roszkowski [39] explores dominant semantic sequences in the disciplinary discourse of judicial opinions. The analysis starts with a grammar pattern in which a noun is followed by a *that*-clause (N *that* pattern) because prior research suggested that nouns in this pattern indicate the epistemic status of the proposition contained in the *that*-clause and such clauses are central to disciplinary epistemology. One of the major goals of this study was to determine whether judges use this linguistic feature to signal the status of propositions in their argumentation. Equally

Discussion Publishers edition DPR-2026-1332

important was the methodological goal to find out whether the semantic sequence approach could be used to characterize judicial opinions in terms of what is often said in this specific type of legal discourse. The findings document how judicial opinions employ extensively a range of nouns found in the N *that* pattern to perform five main functions: *evaluation, cause, result, confirmation and existence*. For example, the analysis of the expression *the idea that* resulted in the following semantic sequence:

- argument/theory + basis + *the idea that*

indicating that an idea may provide grounds for formulating arguments in legal reasoning as in the example (emphasis added):

At base, the *theory* of transferring a defendant's intent *is premised on the idea that* one engaged in a dangerous felony should understand the risk that the victim of the felony could be killed, even by a confederate. (Miller v. Alabama), [39: p. 143].

Apart from signalling genre or disciplinary specificity, semantic sequences represent an interesting methodological construct. The N *that* pattern is a product of a corpus-based analysis and it is useful as a starting point for a detailed investigation. Semantic sequences reflect the consistency of function manifested through diverse language forms raising the question of frequency. Semantic sequences are very frequent cumulatively, that is, if we consider all the occurrences of individual linguistic realizations of a particular function. Each individual realization in the form of a particular phrase may be relatively infrequent. For example, each of the different phrases used to express evaluation and linked to *argument that* (e.g. *the dissent resorts to the last-ditch argument that… This Court finds unpersuasive the argument that…The United States' argument that… is an astounding assertion The Government of the United States has a valid legal argument that…*) appears rather infrequently (only once or twice in the corpus) but the cumulative frequency of the different types exceeds the frequency of its tokens. There are as many as sixty-five different expressions evaluating the phrase *argument that*.

This methodology of semantic sequence promises a more flexible and refined way of analyzing recurrent patterns,especially in legal genres, such as legal justification, which tend to rely on idiosyncratic and less formulaic language.

3.2 Variation in Legal Discourse

As can be seen in the previous section on phraseology, one of the major issues in research on legal discourse has always been concerned with variation. Tiersma in his well-known book *Legal Language* acknowledges the existence of variation in legal language by noting that "[t]here is great variation in legal language, depending on geographical location, degree of formality, speaking versus writing, and related factors. The language and style of lawyers also differs substantially from one genre of writing to another" [101: p.139]. Prior to the advent of corpus linguistics in legal discourse, scholars had to rely on their own intuition and limited language data to substantiate any claims regarding the ways in which legal discourse could vary.

Discussion Publishers edition DPR-2026-1332

Inevitably, this resulted in partial descriptions of legal discourse relying on incomplete and arbitrarily selected linguistic features.

Biel [14: pp. 4–5] usefully points out how variation in legal discourse can be studied at four basic levels: *externally,* investigating how legal language differs from other specialized languages and general language; *internally*, by making explicit comparisons between different legal text types or genres; *diachronically*, analyzing how legal language in its present-day form differs from a historical one and *cross-linguistically* to find out how legal language varies across different languages. These general dimensions of variation have been explored to a varying extent in corpus research depending on a particular approach adopted and the availability of data. Not surprisingly, the study of spoken legal discourse in earlier periods is hampered by the scarcity of authentic data. A recently edited volume [31] has corroborated the relevance of these dimensions in current corpus-based research on variation in English legal discourse by focusing on two dimensions where variation can be fruitfully examined: *cross-genre* (e.g. [15] *and cross-linguistic variation* (e.g. [29] and *diachronic variation* [48].

Historically, research on variation in legal discourse progressed from *textualization* of lexico-grammar through *organization* of discourse to *contextualization* of discourse reflecting a wider trend in the historical development of specialized discourse analysis (cf. [7: p. 4]). Yet, corpus research seems to move in a somewhat non-linear way, with studies displaying characteristics from different stages of the research.

Typical of the textualization stage is early research dating back to the 1960s and 1970s (e.g. [28, 54, 55] which focused on statistically significant features of lexico-grammar used in a particular type of legal texts. Gustaffson [55] investigated the single feature of binomials and multinomials in legal discourse. Almost forty years later, a comparative study of noun binomials (e.g. *terms and conditions, rights and liabilities*) relies on two corpora of UK and Scottish legislation exceeding 14 million words and enriched with an automatic semantic analysis using the USAS online tool [65]. For all its impressive scope and, undoubtedly, the breadth of incorporating semantic motivations for the use of noun binomials, the study adopts a similar approach by taking a single linguistic feature as its main focus of analysis.

The concern with the organization stage was mainly due to the theoretical framework known now as ‘move analysis’ [98]. Its aim was to identify prevalent patterns of discourse organization in a given genre. Although, initially, qualitative and based on a limited range of texts, move analysis research has been recently developed to adopt a corpus-based, quantitative perspective [9, 48]. It is the contextualisation stage that remains the biggest challenge for corpus research prompting corpus linguistics researchers to resort to a triangulation approach.

The recognition that legal discourse is extremely complex and it is embedded in highly varied institutional space of different legal systems and cultures led to a number of studies which aimed at proposing various taxonomies and typologies [5, 101, 106] in an attempt to provide more accurate and meaningful descriptions of legal language. For example, Tiersma [101] proposes a general division of legal texts into three major categories of *operative* legal documents (those that create or modify legal relations such petitions, statutes, contracts, wills, etc.), *expository* documents

Discussion Publishers edition DPR-2026-1332

(e.g. judicial opinions which analyse objectively legal points and *persuasive* documents (e.g. briefs or memoranda). A more detailed classification, proposed by Bhatia [5], acknowledges the basic distinction into written and spoken modes and it lists several functional categories (e.g. law reports, cases, judgments, lawyer-client interaction) based on various criteria such as their communicative purpose or the settings in which they are found. The existence of variation in legal language was acknowledged but such claims could not be substantiated and explored in a systematic manner. Despite the recognition that legal language is highly diverse, many linguistic studies have treated it as monolithic [2, 28, 78] best described in terms of several distinctive lexico-grammatical features such as, for example, excessive use of the passive voice, conditionals, archaic adverbs and prepositional phrases, the use of *shall,* etc.), all of which should apparently hold true for all types and categories of legal texts.

3.2.1 Register and Genre Perspectives on Legal Discourse

Crucial to corpus linguistic investigations of variation in legal language has been the use of the concepts of *register* and *genre* in the construction of analytical frameworks. A comprehensive overview of how the concepts of *register*, *genre* and s*tyle* could be used to research language variation can be found in Biber and Conrad [8].

Biber's corpus-based work on registers has been particularly influential by paving the way for adopting the register and genre perspectives on legal discourse. Biber takes the credit for introducing a much needed rigour into the empirical studies of language. The methodology of Multi-Dimensional analysis [42: pp. 44–51] used a statistical procedure known as *factor analysis* in order to identify sets of linguistic features co-occurring in texts. The patterning of linguistic features in a corpus creates linguistic *dimensions* which correspond to salient functional distinctions within a register or genre and it enables comparison between different language varieties. This methodology was applied to analyze variation in a range of specialised discourse domains. It was used in one of the earliest large-scale corpus explorations of variation in American legal discourse [42] in two ways: (1) to compare written legal genres with other specialised (non-legal) and general genres in English and (2) to compare legal genres with one another; exploring patterns of variation *within* legal discourse. More recently, Yuxiu Sun and Le Cheng [113] apply Multi-Dimensional analysis to explore cross-language variation in legislative discourse based on three corpora (Chinese legislative corpus, the corresponding English translation corpus and American legislative corpus).

It should be pointed out that the MD analysis or, in fact, any corpus analysis, is hardly ever confined to quantitative techniques. Rather, qualitative analysis is needed to interpret the functional bases underlying each set of co-occurring linguistic features. The dimensions of variation reflect both linguistic and functional content. Co-occurrence or co-selection of linguistic features emanates from shared function(s). This means that co-occurrence patterns need to be interpreted in terms of situational, social and cognitive functions shared by the linguistic features.

Discussion Publishers edition DPR-2026-1332

One of the most advanced recent corpus-based explorations following and extending this research path is an in-depth study of EU judicial language and its impact on the language of national judges [67]. It is, to date, the first large-scale corpus analysis of intra-linguistic variation in judicial discourse. The study examines both the macrostructure and the microstructure (lexico-grammar patterns, formulaicity, terminology) using a mixed genre-register approach to the linguistic profiling of judgments.

3.3 Legal Discourse in Forensic Contexts

One of the earliest applications of Corpus Linguistics was to assist linguists in their role as an expert witness in the courtroom [23, 68]. Known as Forensic Linguistics [24–26, 51], this law-related language study has come to rely on corpora as a useful instrument of investigation, especially regarding issues such as authorship and plagiarism [81: p. 20]. In fact, the emergence of Forensic Linguistics as a discipline is closely related to cases of disputed authorship in police statements in the UK [24]. Its scope is, of course, much broader and many of its areas call for analytical tools other than the Corpus Linguistics methodology, as evidenced by the following (non-exhaustive) list of subdomains: authorship attribution and plagiarism, the language of the police and law enforcement, interviews with children and vulnerable witnesses in the legal system, courtroom interaction, linguistic evidence and expert testimony in courtrooms, forensic phonetics and speaker identification. These may pose a number of challenges for the analyst, which call for the application of new technologies to the analysis of questioned texts [53].

It is instructive to acknowledge the unique nature of a forensic corpus in terms of its construction and sampling. The importance of these issues become immediately clear when we consider a range of text types routinely encountered in forensic linguistics casework. Cotterill [23: p. 579] lists the following examples: threat letters, suicide notes, blackmail/extortion letters, terrorist/bomb threats, ransom demands, e-mails, text messages police and witness statements, plagiarised texts. All the text types have been under-researched in general linguistics leading to serious deficiencies in their generic knowledge. Not surprisingly, forensic linguists are faced with the so-called ‘data problem’ [63] in much the same way as linguists exploring legal discourse in earlier periods. As Cotterill [23: p. 578] explains:

> Unfortunately, and frustratingly for the forensic linguist, any old collection of texts is precisely what is provided by either the police or solicitors, who have trawled the home, office and computer of a suspect for any texts which are available. They are unaware of genre/register differences, variations in text size and temporal factors, all of which may influence the potential of texts to be analysed.

This means that there are a number of constraints which may restrict the use of corpus linguistics in forensic contexts, mainly due to data scarcity, the length of the texts, especially in the case of incriminated or questioned text whose authorship needs to be determined. For example, it is not uncommon for such texts to consist of

Discussion Publishers edition DPR-2026-1332

no more than a few words and, worse still, be completely decontextualised, as their availability may depend on several different factors, such as, for example, the effectiveness of police searches. Suicide notes are a fine example of such texts (see [115] for a description of the current state of the art research on the corpus-based analysis of suicide notes).

It should be borne in mind that the forensic linguist's work is necessarily comparative [68]. For example, a typical case of disputed authorship involves comparing or contrasting the questioned text with a series of texts (or a single text) identified previously as being created by one or more potential authors. Accordingly, this type of analysis is carried out by analysing linguistic features of two sets of texts: the disputed and the known texts. The results may lead the forensic linguist to reach any of the following conclusions: (a) authorship attribution, (b) authorship identification, (c) determination of the degree of similarity between disputed texts and known texts (d) elimination of one or more suspect authors, and (e) neither elimination nor identification because the investigation is inconclusive as not enough linguistic evidence has been found to support either hypothesis (c) [53].

The need for integrating computer-based and language-based approaches is highlighted in the important area of plagiarism detection [50]. Combining specialist plagiarism detection tool, such as CopycatchGoldv2 [111] with"language-based approaches, either form-only-based or integrated, seem to provide more complete language evidence to sustain a real case of plagiarism" [50: p. 119].

This section could be summed by referring to a piece of advice from two seasoned researchers [51]:

> The novice reader in Forensic Linguistics may be disappointed to know that there is no single valid method or recipe to analyse a forensic text. In each case, the forensic linguist must carefully define the speech evento, study the data, make general observations, formulate null and alternative hypotheses, choose the most convenient linguistic and non-linguistic tools to find the trace that every forensic text leaves, analyse the data objectively, systematically, and accurately, and reach conclusions grounded in the findings obtained.

3.4 Evaluative Language in Legal Discourse

It is relatively recently that the evaluative function of language in legal discourse has begun to attract more attention among legal linguists. A recent overview of the concept and roles of evaluation in legal (judicial) discourse can be found in [40]. Much of the ground covered in this section can be found there.This upsurge of interest mirrors a similar trend found in the study of evaluation in general linguistics [1] and the existing research in legal discourse has largely relied on the plethora of terms and concepts found in other domains of language study. Defined simply as "the broad cover term for the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about" [103: p. 5] evaluation has also been explored as evaluative language (e.g. [44, 74]), *stance* (e.g. [34, 36, 56]), *stance-taking* (e.g. [100]) or *appraisal* (e.g. [58, 83]). Related concepts include *interpersonality* [17], *voice* [16] and *modality* (e.g. [18]).

Discussion Publishers edition DPR-2026-1332

The importance of evaluation in legal discourse is self-evident for at least two reasons. First, it brings to light the importance of linguistic interactions between various legal actors, especially viewed from the perspective of stance and stance-taking (see, for example [99] which focuses on speech patterns in courtroom interactions). Second, the concept of evaluation is central to researching the quality and strategy of legal argumentation. While research in Argumentation Theory tends to assess the merits of legal argumentation based on certain norms of rationality, which serve as a basis for establishing whether an argument is sound and rational, linguists are not only interested in identifying recurrent language patterns in the expression of evaluation but they also focus on the complex relationships of (dis)alignment between the speaker/writer, their interactants and the evaluated object.

Evaluation represents an area of difficulty for researchers, especially those who prioritize corpus linguistics as their principal method of investigation [45, 60]. If we assume that evaluation should be associated with a meaning or a type of meaning rather than a form, then any methodological perspective based on identifying and quantifying forms will lead to considerably impoverished findings. Even more daunting is the question of context in which evaluation is analysed. Many corpus-based studies deal with judicial justifications, a type of discourse marked by numerous conventions and institutional constraints leaving seemingly little room for meanings that might be termed as 'subjective' or 'attitudinal'. Recent research shows that subjectivity and attitudinal language are an indispensable part of the judicial voice [58, 75] and evaluative language is used in various assessments made by judges in the context of justifying their decisions. This is particularly evident in constitutional cases with strong axiological concerns underpinning conflicts between norms and values.

3.4.1 Corpus-based Studies of Evaluation in Judicial Discourse

Most studies of evaluation in judicial discourse are corpus-based, i.e. they start with a pre-defined language form already associated with evaluative meanings, which is then identified and investigated in large corpus data. For example, both Mazzi [75] and Finegan [34] examine the use of adverbials of stance in judicial discourse. The former study focuses on eight stance adverbs (e.g. *apparently, clearly*, etc.) analyzed in a corpus of 98 equity judgments of the Chancery Division of the High Court of Justice of England and Wales. In the latter [34] judicial attitude is examined by focusing on adverbial expressions of attitudinal stance and emphasis (e.g. *simply, certainly*). There is a long tradition of investigating stance adverbials in English ever since the classic study conducted by Conrad and Biber [21]. The approach here is largely quantitative and it relies on the frequencies of specific forms (e.g. adverbials) in different datasets to provide descriptive statistics and observations about different registers or genres or even languages. In other words, this approach pre-supposes a comparative perspective. Thus, Finegan [34] can demonstrate the relative frequency of single adverbs in the domain-specific corpus of US Supreme Court opinions as compared with their frequencies in general language corpora.

The limitation of this approach is that it usually isolates a relatively small number of features and the interpretation of the descriptive statistics is confined to the

Discussion Publishers edition DPR-2026-1332

distributive trends in the corpora under investigation. For one, it would be certainly interesting to learn why judges keep making use of a linguistic feature (e.g. *clearly*) that has been criticized and discouraged in numerous handbooks of legal style [37]. But the answer to this question should be sought in the socio-cognitive and professional space [7] rather than in the textual space.

In a similar vein, other corpus-based studies of stance or evaluative language start with a language pattern, either already associated with evaluative language, or with a pattern that is viewed as potentially evaluative. Corpus evidence is used to prove or disprove the evaluative potential of a given language item. For example, a study documented in [44] uses the notions of grammar patterns [62] and local grammar [49] to analyze adjectival patterns in judgments given by the US Supreme Court and the Italian *Corte Suprema di Cassazione*. Rather than describing a language as a whole, corpus grammarians developed the idea that certain areas of language use could be examined more effectively, if treated separately, as they seem to show patterning of their own, which often does not always correspond with the generalised categories postulated in standard grammars (see [61] for a more recent discussion of this issue). The study compares the use of two patterns: *v-link + ADJ + that* and *v-link + ADJ + to-infinitive* with the equivalent patterns in Italian: *copula + ADJ + che* and *copula + ADJ + verbo all'infinito*. These patterns are used as a diagnostic for retrieving instances of evaluative language manifested through the use of adjectives. The advantage of using such linguistic constructs is that they represent a viable unit of analyzing evaluation not only in monolingual corpora but also across different languages.

A variation on this approach are corpus-based investigations undertaken to determine the evaluative potential of a pre-defined language item or pattern. For example, Mazzi [74] focuses on the single discourse element of 'this/these/that/those + the labelling noun' and provides some corpus evidence to demonstrate that abstract nouns such as *attitude, difficulty, process, reason*, etc. have both encapsulating and evaluative function, when found in this pattern in the US Supreme Court judicial opinions. Interestingly, Mazzi [74] selects this pattern after a qualitative analysis, i.e. when a randomly chosen judicial opinion has been manually explored to identify salient expressions of evaluation. A corpus of US opinions is then used to verify the hypothesis that the pre-selected pattern could indeed be associated with evaluative meanings.

There are also large-scale studies which focus on modality as yet another aspect of evaluation. Note that the 'combining' approach to evaluation may include modality [21, 103]. Such studies are usually firmly grounded in a specific theoretical framework in order to gain initial input but they use corpus linguistics to retrieve, analyze and verify the data. For example, Strebska-Liszewska [96] in her research of modality in the opinions of the US and Polish Supreme Courts uses the Systemic Functional Linguistics and the Hallidayan classification of high, median and low-value markers of epistemic modality as the theoretical framework. A range of pre-defined epistemic markers are subsequently interrogated and analyzed using substantial corpus data since the author is primarily interested in the distribution of the epistemic markers across different sections of the investigated material. A similar approach is adopted in [112] which traces the distribution of modal verbs (e.g.

Discussion Publishers edition DPR-2026-1332

may, shall, must) in three sets of corpora (Uniform Commercial Code, United States Code, and the Freiburg-LOB Corpus of American English) based on the theoretical framework of Halliday's systemic-functional grammar.

Most corpus-based studies vary in the degree to which corpus analysis is accompanied or augmented by other types of investigation. It is relatively seldom that research into evaluative language is carried out based on corpus analysis alone [45] For example, Szczyrbak [100] focuses on phrases with 'say' as indicative of the alignment function of stance-taking. The paper identifies the frequency of each phrase in the corpus of libel proceedings before a UK court. It locates each phrase in an extended co-text, identifying a series of distinct pragmatic functions routinely performed by the phrases. Identified explicitly as 'corpus-assisted', this approach effectively illustrates the interaction of a corpus approach that prioritizes rapid searches for specific forms and quantitative comparisons between corpora, and a discourse approach that examines and interprets the discourse surrounding the target item to identify its function. Those functions cannot be effectively extracted from frequency lists or even from limited concordance lines. Yet, it is the corpus input that identified phrases most likely to be significant.

This brief overview demonstrates that corpora can be used effectively to quantify overt expressions of evaluation and corpus linguistics tools can help to determine the evaluative potential of those items that may not be inherently evaluative. Yet, any such work should be complemented by a more qualitative approach. A corpus-assisted discourse analysis could be preferable because it treats descriptive statistics as the starting point for further explorations rather than as their end. This is particularly important in the case of such context-sensitive and pragmatic phenomenon as evaluation. An act of evaluation is never performed in a vacuum. When evaluation is enacted in a professional space of law, it is underpinned by the tension between the legal actor's own individual position and a position which reflects the epistemological beliefs, values and constraints of a given legal community.

When explored in the context of legal argumentation, it seems that linguistically-oriented studies of evaluation are limited in that they often fail to account for the macro-context of the institutional environment in which argumentative discourse occurs. This means that the results of linguistic investigations are not grounded in the argumentative reality of judicial discourse. One possible solution is to integrate the analysis at the microlevel of discrete language items with a theory of legal argumentation [33].

4 Summary and Conclusions

The enormous expansion of Corpus Linguistics and the availability of digital texts and computer tools have paved the way for large-scale analyses investigating a wide range of linguistic features based on large amounts of multi-genre data that were simply not feasible earlier. Corpus studies of legal discourse have revealed patterns of use and distribution that are not visible when examining individual texts or even many texts individually. They have brought greater scientific rigour thanks to

Discussion Publishers edition DPR-2026-1332

systematic explorations of reliable, representative and balanced corpora, superseding anecdotal illustrations based on scant and random textual evidence.

Some of the criticism levelled at using corpus linguistics in legal discourse [6] relates to the early stage of quantifying lexico-grammatical features in single (usually legislative) genres, where the form-function correlations are indeed highly formulaic rendering corpus analysis somewhat superfluous. Thanks to the register and genre approach, the empirical turn in corpus studies of legal discourse has shifted from surface-level analyses of lexico-grammatical forms focusing on textual patterns and confined to textual space to deeper levels of socio-cognitive professional contexts extending to the pragmatic: the tactical and social space [38].

More serious limitations can be traced to corpus research being inevitably impoverished in the sense that methodologies based on searching for and quantifying specific features rely on those features being identifiable without human intervention. Quantitative corpus methods turn out to be inadequate when dealing with meaning-based and context-sensitive phenomena such as evaluation, which need not be attached to any specific language resources.

Many corpus explorations of legal discourse adopt the designation 'legal language' rather than 'legal discourse'. This is not surprising considering that the term 'discourse' is usually reserved for naturally-occurring written or spoken language produced in institutional and professional contexts [7]. It is precisely the non-linguistic, institutional and professional contexts of law that have exposed the most serious deficiencies in corpus methodologies. Nowhere is that more conspicuous than in the study of legal argumentation, where the macro-context of the institutional environment in which argumentative discourse occurs is crucial to its proper interpretation and understanding. This issue has been only partly remedied by resorting to qualitative analytical frameworks, such as the Pragma-Dialectical Theory of Argumentation or Critical Discourse Analysis (CDA) and adding sources of information other than the corpus. It should have become now clear that Corpus Linguistics might not always offer solutions to research questions in the study of legal discourse. For example, many pragmatic features such as implicit speech acts, politeness, hedges, boosters, vague language, and so on, are not automatically retrievable from a corpus. As a result, triangulation of data sources and methods are becoming the trademark of contemporary Corpus Linguistics.

This article has reviewed some of the core linguistic applications focusing on the central idea of co-occurrence and co-selection in language construction approached from the phraseological perspective on language. It has not addressed the so-called 'legal applications' of using Corpus Linguistics methods. There are several reasons for this. First, there have already been some attempts made by lawyer linguists, best positioned, to assess and present this strand of corpus research [91, 93, 107]. Second, it seems that Corpus Linguistics is often used as a pliant umbrella term for almost any computer-supported analysis of collections of legal texts. While many such studies prove extremely valuable, they have little to do with linguistics. One very recent example is [71] which analyzes argumentation strategies (in terms of topoi, standard figures of argumentation that are widely accepted in the respective legal cultures) applied by the Polish Supreme Court and the German Federal Court of Justice judges in their decisions. The description of the dataset does not include

Discussion Publishers edition DPR-2026-1332

any rationale for the sampling frame, balance and representativeness routinely found in Corpus Linguistics studies. Third, some legal applications, such as citation analysis [107: p. 1348] are irrelevant for linguists (and the same is certainly true of many linguistic applications in the legal eyes). Still others appear local and confined to a specific jurisdiction, as in the case of using corpora as a tool in legal interpretation in the United States common law system, to determine which meanings are ordinary [91]. This does not mean that linguistic and legal applications cannot overlap and lead to cross-fertilization and richer understanding of legal discourse. But as Vogel et al. [107: p. 1357] rightly point out such endeavours would call for a mutual cooperation among linguists, lawyers and computational scholars, who would need to agree on a"a common (meta) language as well as an intercultural understanding of the interests, basic theoretical backgrounds, methods, and limitations of each of these disciplines". It seems we are still a long way from being able to attain this goal.

Finally, it is sometimes said that Corpus linguistics is a discipline where technological advancement and theoretical development go hand in hand [60]. The research infrastructure offering data, tools and services to support research based on language resources is developing at a mind-boggling rate. It remains to be seen to what extent research into legal discourse will continue to draw on resources and theoretical insights afforded by Corpus Linguistics adapting them to the unique nature and settings of legal texts.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Alba-Juez, Laura, and Geoff Thompson. 2015. The many faces of evaluation. In *Evaluation in context*, ed. Geoff Thompson and Laura Alba-Juez, 3–23. Amsterdam: John Benjamins.
2. Alcaraz, Enrique, and Brian Hughes. 2002. *Legal translation explained*. Manchester: St.Jerome Publishing.
3. Anthony, Laurence. AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software. Accessed 30 June. 2021.
4. Basaneže, K. 2015. Investigating concgrams in the language of contracts and legal agreements. *Fachsprache* 3–4: 176–192.
5. Bhatia, Vijay K. 1993. *Analysing genre. Language use in professional settings*. London: Longman.
6. Bhatia, Vijay, Nicola M. Langton, and Jane Lung. 2004. Legal discourse: Opportunities and threats for corpus linguistics. In *Discourse in the profession. Perspectives from corpus linguistics*, ed. Ulla Connor and Thomas A. Upton, 203–234. Amsterdam: John Benjamins.
7. Bhatia, Vijay. 2004. *Worlds of written discourse. Genre-based View*. Canada: Continuum.

Discussion Publishers edition DPR-2026-1332

8. Biber, Douglas, and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.
9. Biber, Douglas, Connor Ulla, and Thomas Upton. 2007. *Discourse on the move. Using corpus analysis to describe discourse structure*. Amsterdam/Philadelphia: John Benjamins.
10. Biber, D., and F. Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purpose* 26 (3): 263–286.
11. Biel, Łucja. 2018. Corpora in institutional legal translation: Small steps and the big picture. In *Institutional translation for international governance: Enhancing quality in multilingual legal communication*, ed. F. Prieto Ramos, 25–36. London: Bloomsbury.
12. Biel, Łucja. 2018. Lexical bundles in EU law: The impact of translation process on the patterning of legal language. In *Phraseology in legal and institutional settings. A corpus-based interdisciplinary perspective*, ed. Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo, 11–26. London: Routledge.
13. Biel, Łucja. 2014. *Lost in the Eurofog. The textual fit of translated law*. Frankfurt am Main: Peter Lang.
14. Biel, Łucja. 2010. Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009,*ed. Heine, Carmen and Engberg. Aarhus.
15. Breeze, Ruth. 2019. Part-of-speech patterns in legal genres: Text-internal dynamics from a corpus-based perspective. In *Corpus-based research on variation in English legal discourse*, ed. Teresa Fanego and Paula Rodriguez-Puente, 79–103. Amsterdam: John Benjamins.
16. Breeze, Ruth. 2018. Giving voice to the law. Speech act verbs in legal academic writing. In *Phraseology in legal and institutional settings. A corpus-based interdisciplinary perspective*, ed. Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo, 221–239. London: Routledge.
17. Breeze, Ruth, Maurizio Gotti, Sancho Guinda, and Carmen , eds. 2014. *Interpersonality in legal genres*. Frankfurt: Peter Lang.
18. Cheng, Winnie, and Le. Cheng. 2014. Epistemic modality in court judgments: A corpus-driven comparison of civil cases in Hong Kong and Scotland. *English for Specific Purposes* 33: 15–26.
19. Cheng, Winnie, Chris Greaves, and Martin Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11 (4): 411–433. https://doi.org/10.1075/ijcl.11.4.04che.
20. CLARIN The research infrastructure for language as social and cultural data, https://www.clarin.eu/.
21. Conrad, Susan, and Douglas Biber. 2000. Adverbial marking of stance in speech and writing. In *Evaluation in text authorial stance and the construction of discourse*, ed. Geoff Thompson and Susan Hunston, 56–73. Oxford: Oxford University Press.
22. Corpus of US Supreme Court Opinions, https://www.english-corpora.org/scotus/.
23. Cotterill, Janet. 2010. How to use corpus linguistics in forensic linguistics. In *The Routledge handbook of corpus linguistics*, ed. Anne O'Keefe and Michael McCarthy, 578–590. London: Routledge.
24. Coulthard, Malcolm, Alison Johnson, and David Wright, eds. 2017. *An introduction to forensic linguistics: Language in evidence*, 2nd ed. London: Routledge.
25. Coulthard, Malcolm, and Johnson Alison, eds. 2010. *The Routledge handbook of forensic linguistics*. London & New York: Routledge.
26. Coulthard, Malcolm. 1994. On the use of corpora in the analysis of forensic texts. *The International Journal of Speech, Language and the Law* 1 (1): 27–43.
27. Cowie, A.P. 1994. Phraseology. In *The encyclopedia of language and linguistics*, ed. R.E. Asher, 3168–3171. Oxford: Oxford University Press.
28. Crystal, David, and Derek Davy. 1969. *Investigating English style*. London: Longman.
29. Diani, Giuliana. 2019. English and Italian land contracts: A cross-linguistic analysis. In *Corpus-based research on variation in english legal discourse*, ed. Teresa Fanego and Paula Rodriguez-Puente, 25–50. Amsterdam: John Benjamins.
30. Fairclough, N., and R. Wodak. 1997. Critical discourse analysis. In *Discourse studies: A multidisciplinary introduction*, vol. 2, ed. T.A. van Dijk. London: Sage.
31. Fanego, Teresa, and Paula Rodriguez-Puente. 2019. *Corpus-based research on variation in English legal discourse*. Amsterdam: John Benjamins.

Discussion Publishers edition DPR-2026-1332

32. Prieto Ramos, F. 2020. Translating legal terminology and phraseology: Between inter-systemic incongruity and multilingual harmonization. *Perspectives*. https://doi.org/10.1080/0907676X.2021.1849940.
33. Feteris, Eveline. 2017. *Fundamentals of legal argumentation. A survey of theories on the justification of judicial decisions*. London: Springer.
34. Finegan, Edward. 2010. Corpus linguistics approaches to 'legal language': Adverbial expression of attitude and emphasis in Supreme Court opinions. In: Malcolm Coulthard &
35. Fletcher, William. 2002. *KfNgram*. Annapolis MD: USNA.
36. Gales, T. 2015. The stance of stalking: A corpus-based analysis of grammatical markers of stance in threatening communications. Special issue on forensic linguistics. *Corpora* 10 (2): 171–200.
37. Garner, Bryan. 1995. *A dictionary of modern legal usage*. Oxford and New York: Oxford University Press.
38. Giltrow, Janet, and Dieter Stein. 2017. *The pragmatic turn in law. Inference and interpretation in legal discourse*. Boston/Berlin: De Gruyter Mouton.
39. Goźdź-Roszkowski, Stanisław. 2018. Between corpus-based and corpus-driven approaches to textual recurrence. Exploring semantic sequences in judicial discourse. In *Patterns in text: Corpus-driven methods and applications*, ed. Joanna Kopaczyk and Tyrkko Jukka, 131–158. Amsterdam: John Benjamins.
40. Goźdź-Roszkowski, Stanisław. 2018. Counting the uncountable? Quantitative and qualitative methods of analyzing evaluative language in institutional discourse. A corpus linguistic perspective. In *Explorations in language and linguistics*, ed. Łukasz Bogucki and Cap Piotr, 145–158. Frankfurt am Main: Peter Lang.
41. Goźdź-Roszkowski, Stanisław, and G. Pontrandolfo. 2018. Cross-linguistic approaches and applications to phraseology in legal and institutional discourse. In *Phraseology in legal and institutional settings. A corpus-based interdisciplinary perspective*, 1–8. London: Routledge.
42. Goźdź-Roszkowski, Stanisław. 2011. *Patterns of linguistic variation in American legal English. A corpus-based study*. Frankfurt am Main: Peter Lang.
43. Goźdź-Roszkowski, S., and G. Pontrandolfo. 2015. Legal phraseology today. A corpus-based view. *Fachsprache* 3–4: 130–138.
44. Goźdź-Roszkowski, S., and G. Pontrandolfo. 2014. Exploring the local grammar of evaluation: The case of adjectival patterns in American and Italian judicial discourse. *Research in Language* 12 (1): 71–92.
45. Goźdź-Roszkowski, S., and S. Hunston. 2016. Corpora and beyond – investigating evaluation in discourse: Introduction to the special issue on corpus approaches to evaluation. *Corpora* 11 (2): 131–141. https://doi.org/10.3366/cor.2016.0089.
46. Granger, Sylviane, and Magali Paquot, eds. 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.
47. Gries, Stefan. 2008. Phraseology and linguistics theory. In *Phraseology: An interdisciplinary perspective*, ed. Sylviane Granger and Fanny Meunier, 3–25. Amsterdam: John Benjamins.
48. Groom, Nicholas, and Jack Grieve. 2019. The evolution of a legal genre. Rhetorical moves in British patent specifications 1711 to 1860. In *Corpus-based research on variation in English legal discourse*, ed. Teresa Fanego and Paula Rodriguez-Puente, 201–234. Amsterdam: John Benjamins.
49. Gross, Maurice. 1993. Local grammars and their representation by finite automata. In *Data, description, discourse: Papers on the English language in Honour of John McH Sinclair*, ed. Michael Hoey, 26–38. New York: HarperCollins Publishers.
50. Guillén-Nieto, Victoria. 2020. The relevance of context in plagiarism detection: The case of a professional legal genre. *Ibérica* 40 (2020): 101–122.
51. Guillén-Nieto, Victoria, and Dieter Stein. 2021. Introduction: Theory and practice in forensic linguistics. In *Language as evidence: Doing forensic linguistics*, ed. Victora Guillén-Nieto and Dieter Stein. London: Palgrave.
52. Guillén-Nieto, Victora, and Dieter Stein, eds. 2021. *Language as evidence: Doing forensic linguistics*. London: Palgrave.
53. Guillén-Nieto, V., C. Vargas-Sierra, M. Pardiño-Juan, P. Martínez-Barco, and A. Suárez-Cueto. 2008. Exploring state-of-the-art software for forensic authorship identification. *International Journal of English Studies. IJES* 8 (1): 1–28.
54. Gustafsson, M. 1984. The syntactic features of binomial expressions in legal English. *Text Interdisciplinary Jounal for the Study of Discourse* 4 (123): 132.

Discussion Publishers edition DPR-2026-1332

55. Gustafsson, Maria. 1975. *Some syntatic properties of English law language. Publications of the department of English 4*. Turku: University of Turku.
56. Hafner, Christoph. 2014. Stance in a professional legal genre: The barrister's opinion. In *Interpersonality in legal genres*, ed. Ruth Breeze, Maurizio Gotti, and Carmen Sancho Guinda, 137–160. Bern: Peter Lang.
57. Hajlaoui, Najeh, Kolovratnik, David, Vaeyrynen, Jaakko, Steinberger Ralf, and Varga, Dániel. 2014. DCEP-Digital Corpus of the European Parliament. In: Proc. LREC 2014 (Language Resources and Evaluation Conference). Reykjavik, Iceland. Mai pp. 26–31, 2014. pp 3164–3171 (URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/943_Paper.pdf).
58. Heffer, Christopher. 2007. Judgement in court: Evaluating participants in courtroom discourse. In *Language and the law: International outlooks*, ed. Krzysztof Kredens and Stanisław Goźdź-Roszkowski, 145–179. Frankfurt am Mein: Peter Lang GmbH.
59. Hoffmann, Sebastian, Bettina Fischer-Starcke, and Sand Andrea, eds. 2015. *Current issues in phraseology*. Amsterdam/Philadelphia: Benjamins.
60. Hunston, Susan. 2011. *Corpus approaches to evaluation. Phraseology and evaluative language*. New York/London: Routledge.
61. Hunston, Susan, and Su. Hang. 2017. Patterns, constructions, and local grammar: A case study of 'evaluation.' *Applied Linguistics*. https://doi.org/10.1093/applin/amx046.
62. Hunston, Susan, and Gill Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
63. Jucker, Andreas H., ed. 1995. *Historical pragmatics. Pragmatic developments in the history of English*. Amsterdam: John Benjamins.
64. Kjaer Anne Lise. 2007. Phrasemes in legal texts. In *Phraseologie: An international handbook of contemporary research,* ed. Harald Burger, Dmitri Dobrovol'skij, Peter Kühn, and Neal Norrick, pp. 506–516. Berlin: de Gruyter.
65. Kopaczyk, Joanna. 2018. Terms and conditions: A comparative study of noun binomials in UK and Scottish legislation. In *Phraseology in legal and institutional settings. A corpus-based interdisciplinary perspective*, ed. Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo, 160–186. London: Routledge.
66. Kopaczyk, Joanna. 2013. *The legal language of Scottish Burghs. Standardization and lexical bundles*. Oxford: Oxford University Press.
67. Koźbiał, Dariusz. 2020. *The language of EU and polish judges. Investigating textual fit through corpus methods*. Bern: Peter Lang.
68. Kredens, Krzysztof, and Malcolm Coulthard. 2012. Corpus linguistics in authorship identification. In *The Oxford handbook of language and law*, ed. Peter Tiersma and Lawrence Solan. Oxford: Oxford University Press.
69. Law and Language at the European Court of Justice Project, https://llecj.karenmcauliffe.com/.
70. Leung, E.S.M. 2014. What can a bilingual corpus tell us about the translation and interpretation of rape trials? *International Journal for the Semiotics of Law* 28: 469–483. https://doi.org/10.1007/s11196-014-9400-y.
71. Małolepszy, M., and M. Głuchowski. 2021. Argumentation and legal interpretation in the criminal decisions of the polish supreme court and the German federal court of justice: A comparative view. *International Journal for the Semiotics of Law*. https://doi.org/10.1007/s11196-021-09843-9.
72. Pérez, M.J.M., and C.R. Rizzo. 2012. Structure and design of the British law report corpus (BLRC): A legal corpus of judicial decisions from the UK. *Journal of English Studies* 10: 131–145.
73. Mazzi, Davide. 2022. Without proof of negligence or a causative connection": On causal argumentation in supreme court of Ireland's judgments on data protection. In *Law, language and the courtroom legal linguistics and the discourse of judges*, ed. Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo. London: Routledge.
74. Mazzi, D. 2010. This argument fails for two reasons. A linguistic analysis of judicial evaluation strategies in US supreme court judgements. *International Journal for the Semiotics of Law* 23 (4): 373–385.
75. Mazzi, D. 2007. The construction of argumentation in judicial texts: Combining a genre and a corpus perspective. *Argumentation* 21 (1): 21–38.
76. McCarthy, Michael, and Anne O'Keeffe. 2010. Historical perspective what are corpora and how have they evolved? In *The Routledge handbook of corpus linguistics*, ed. Anne O'Keefe and Michael McCarthy, 3–13. London: Routledge.

Discussion Publishers edition DPR-2026-1332

77. McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies. An advanced resource book*. London and New York: Routledge.
78. Mellinkoff, David. 1963. *Language of the law*. Boston, MA: Little Brown.
79. Monzó Nebot, Esther. 2018. The out-grouping society: phrasemes othering underprivileged groups in the international bill of human rights (English-French-Spanish). In *Phraseology in legal and institutional settings. A corpus-based interdisciplinary perspective*, ed. Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo, 109–125. London: Routledge.
80. O'Keefe, Anne, and Michael McCarthy. 2010. *The Routledge handbook of corpus linguistics*. London: Routledge.
81. O'Keeffe, Anne, Michale McCarthy, and Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
82. Partington, Alan, Duguid Alison, and Charlotte Taylor. 2013. *Patterns and meanings in discourse. Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam: John Benjamins.
83. Pérez, María José Marín. 2022. A corpus-based comparative analysis of the evaluative lexicon found in judicial decisions on immigration. In *Law, language and the courtroom legal linguistics and the discourse of judges*, ed. Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo. London: Routledge.
84. Pęzik, Piotr. 2018. *Facets of prefabrication. Perspectives on modelling and detecting phraseological units*. Łódź: Lodz University Press.
85. Pontrandolfo, Gianluca. 2019. Corpus methods in legal translation studies. In *Research methods in legal translation and interpreting crossing methodological boundaries,* ed. Łucja Biel, Jan Engberg, M. Rosario Martín Ruano, and Vilelmini Sosoni, 13-28. London and New York: Routledge.
86. Pontrandolfo, Gianluca. 2012. Legal corpora: An overview. In ed. M. Magris & H. Lozano *RITT (Rivista Internazionale di Tecnica della Traduzione)*, 14: 121–136, Trieste: EUT.
87. Potts, A., and A.L. Kjær. 2016. Constructing achievement in the international criminal tribunal for the former Yugoslavia (ICTY): A corpus-based critical discourse analysis. *International Journal for the Semiotics of Law* 29: 525–555. https://doi.org/10.1007/s11196-015-9440-y.
88. Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13 (4): 519–549. https://doi.org/10.1075/ijcl.13.4.06ray.
89. Ruusila, A., and E. Lindroos. 2016. Conditio sine qua non. On phraseology in legal language and its translation. *Language and Law/Linguagem e Direito* 3 (1): 120–140.
90. Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
91. Solan, L.M. 2020. Corpus linguistics as a method of legal interpretation: Some progress, some questions. *International Journals for the Semiotics of Law* 33: 283–298. https://doi.org/10.1007/s11196-020-09707-8.
92. Solan, Lawrence. 1993. *The language of judges*. Chicago: The University of Chicago Press.
93. Solan, L., and T. Gales. 2017. corpus linguistics as a tool in legal interpretation. *Brigham Young University Law Review* 6: 1311–1358.
94. SOULL (Sources of Language and Law), https://legal-linguistics.net/data-collections.
95. Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, Dániel Varga. 2006. [application/pdf] The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, pp. 24–26.
96. Strębska-Liszewska Katarzyna. 2017. Epistemic modality in the rulings of the American Supreme Court and Polish Sąd Najwyższy: A corpus-based analysis of judicial discourse. (unpublished Phd thesis) Uniwersytet Śląski.
97. Swales, John. 2004. *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.
98. Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
99. Szczyrbak, Magdalena. 2018. *Verba dicendi* in courtroom interaction: Patterns with the progressive. In *Phraseology in legal and institutional settings. A corpus-based interdisciplinary perspective*, ed. Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo, 240–257. London: Routledge.
100. Szczyrbak, M. 2016. Say and stancetaking in courtroom talk: A corpus-based study. *Corpora* 11 (2): 143–168.
101. Tiersma, Peter. 1999. *Legal language*. Chicago University Press.
102. Tiersma, Peter, and Lawrence Solan, eds. 2012. *The Oxford handbook of language and law*. Oxford: Oxford University Press.

Discussion Publishers edition DPR-2026-1332

103. Thompson, Geoff, and Susan Hunston. 2000. Evaluation. An introduction. In *Evaluation in text authorial stance and the construction of discourse*, ed. Susan Hunston and Geoff Thompson, 1–27. Oxford: Oxford University Press.
104. Tognini-Bonelli, Elena. 2010. Theoretical overview of the evolution of corpus linguistics. In *The routledge handbook of corpus linguistics*, ed. Anne O'Keefe and Michael McCarthy, 14–27. London: Routledge.
105. Tognini Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam and Philadelphia: John Benjamins.
106. Trosborg, A. 1995. Statutes and contracts: An analysis of legal speech acts in the English language of the law. *Journal of Pragmatics* 23: 31–53.
107. Vogel, F., H. Hamann, and I. Gauer. 2018. Computer-assisted legal linguistics: Corpus analysis as a new tool for legal studies. *Law & Social Inquiry* 43 (4): 1340–1363.
108. Warren, Martin. 2010. Determining aboutgrams in engineering texts. In *Keyness in texts [studies in corpus linguistics 41]*, ed. Marina Bondi and Mike Scott, 113–126. Amsterdam: John Benjamins.
109. Wei, Yu. 2020. Reporting verbs in court judgments of the common law system: A Corpus-Based Study. *International Journal for the Semiotics of Law* 34: 525–560. https://doi.org/10.1007/s11196-020-09740-7.
110. Whittaker, S. 2014. Contrastive legal discourse research. Corpus design and applications. *Revue française de linguistique appliquée* 1 (1): 49–57. https://doi.org/10.3917/rfla.191.0049.
111. Woolls, David. 2002. *Copycatch gold v2*. United Kingdom: CFL Software Development.
112. Xinyu, Wu., and Jian Li. 2019. A corpus-based study of modal verbs in the uniform commercial code of the USA. *International Journal for the Semiotics of Law* 32: 463–483.
113. Sun, Y., and Le. Cheng. 2017. Linguistic variation and legal representation in legislative discourse: A corpus-based multi-dimensional study. *International Journal of Legal Discourse* 2: 315–339. https://doi.org/10.1515/ijld-2017-0017.
114. Zanettin, Federico. 2012. *Translation-driven corpora. Corpus resources for descriptive and applied translation studies*. Manchester: St. Jerome Publishing.
115. Zaśko-Zielińska, M. 2021. The linguistic analysis of suicide notes. In: *Language As Evidence: Doing Forensic Linguistics,* ed. Guillén-Nieto, Victora & Dieter Stein. Palgrave.
116. Ziemski, M, Junczys-Dowmunt, M, and Pouliquen, B. 2016. The United Nations parallel corpus, language resources and evaluation (LREC'16), Portorož, Slovenia.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Discussion Publishers edition DPR-2026-1332