Legal Phraseology Today: Corpus-based Applications Across Legal Languages and Genres

Stanisław Goźdź-Roszkowski & Gianluca Pontrandolfo

1 Introduction

One of the main theoretical findings of recent linguistics is that phraseology in the broad sense is central to language organisation (Hoffmann et al. 2015). Phraseological items may come in different shapes and sizes but they take precedence over single words. Yet, for all the impressive growth of studies into “[the] structure, meaning and use of word combinations” (Cowie 1994: 3168) over the past thirty years, research into phraseology in specialist discourse domains tended to lag behind the developments in NLP (Natural Language Processing) (Schulze/Römer 2008). There have been several reasons often cited to account for the perceived gap between LSP (Language for Specific Purpose) and LGP (Language for General Purposes) phraseology studies. Undoubtedly, many problems can be traced to the absence of rigorous and replicable methodologies to identify phraseologisms, the lack of clear and explicit criteria for distinguishing between phraseological and non-phraseological units (as well as classifying different types of phraseologisms), the preoccupation with the terminological perspective at the expense of other aspects of LSP phraseology, the scarcity of freely available corpus resources to carry out new analyses and to replicate the existing ones, and finally, the domain-specificity of different disciplinary discourses (legal, medical, scientific, etc.) precluding a more universalist view of phraseology. This list is not exhaustive and it seems that each of these factors, to a varying extent, has contributed to relegating phraseology to the periphery of LSP studies (cf. Kjær 2007: 506).

The legal domain and its phraseology has also received relatively scant attention. Textual recurrence was initially studied in terms of formulaicity traditionally regarded as one of the most typical and conspicuous features of legal style (Crystal/Davy 1969: 194). Formulaic expressions have been found to lie at the core of much of the formal and ritualistic language so ubiquitous in legal proceedings and documents (Tiersma 1999: 100–104). On the one hand, these expressions have been perceived as an obstacle to the understanding of professional-lay communication, especially in the courtroom, but, on the other hand, formulaic expressions have been recognised as vital signposts helping interactants to orient themselves at different stages in the course of legal proceedings (Coulthard/Johnson 2009). One type of formulaic expressions which has been most extensively studied is known as lexical doubling or binomials (e.g. Gustaffson 1984) and shown to be a distinctive feature of legal language. Gustaffson’s analysis demonstrated that such sequences of two or three words (trinomials) belonging to the same part of speech and connected by some lexical link and semantic unity are used five times as often in legal texts as in other prose styles. While the statistics could vary depending on a particular text genre, this study is one of the earliest quantitative examinations of multi-word units in legal language and it is representative of an early interest in exploring the functions of repetitive and fixed expressions in legal discourse. Furthermore, this line of research signalled

---

1 This paper stems from the ideas of both authors. Sections 1, 2 and 5 were written by Stanisław Goźdź-Roszkowski, whereas sections 3 and 4 by Gianluca Pontrandolfo.
the potential for relying on repetition, fixedness and frequency to uncover patterns of lexical
combinations which may be otherwise difficult to intuit.

The crucial link between preserving formulaic patterning in legal instruments and achiev-
ing the corresponding legal effects was explicitly acknowledged in Kjær (1990: 28), who in her
semina paper “Context-conditioned word combinations in legal language” argued that failure
to employ prefabricated word combinations which are directly prescribed by law “[will] result
in the invalidation (our emphasis) of the whole text of which they form a part” (1990: 28).
Even twenty-five years after its publication, this article is noteworthy for at least two reasons.
First, Kjær posed several fundamental questions some of which appear to be relevant for legal
phraseology even today. These questions range from the very notion of phraseology and the
need for its precise definition, the repertoire of different word combinations found in legal
texts, the distinction between LGP and LSP phraseology, and consequently the extent to which
methods used to describe LGP phraseology may be used in the description of legal phraseolo-
y. As can be seen in Section 2 below, contemporary phraseology is still struggling to explicate
its nature and to define parameters of phraseologisms (Gries 2008: 3). The cross-fertilisation
between Corpus Linguistics, phraseology and LSP studies, while promising and innovative,
raises the perennial question of selecting optimum methodologies for investigating word com-
binaitons in specialist discourse domains. It seems that all too often LSP researchers tend to ap-
ply corpus methodologies rather indiscriminately. They are preoccupied with the extraction of
multi-word units without giving adequate attention to their analysis and interpretation in light
of the complexities and specificity of the wider non-linguistic professional contexts in which
the specialist domains are embedded. In fact, the other important point made in Kjær (1990)
is that the explanation for the occurrence of phraseology in legal texts should be sought in the
world of law, rather than within the realm of linguistics. This was (and perhaps still is) a valid
criticism of those analytical approaches that only examine texts generated by legal systems and
their institutions treating them as an artefact detached from the world of reality. But this study
is also noteworthy for its limitations, characteristic of the early period when Corpus Linguistics
was only beginning to weave its way into mainstream linguistics, electronic corpora of legal
texts were virtually non-existent and general linguists were engaged in what Gries (2008: 3)
describes as “idiom researchers and lexicographers classifying and researching various kinds of
fairly frozen idiomatic expressions”. As a result, the scope of these studies was heavily restrict-
ed in that phraseology was mainly studied monolingually as a lexical environment for terms
identified manually within a very narrow range of legal genres (legislative texts being the pre-
ferred genre). It is the advent of Corpus Linguistics that has radically changed the landscape of
contemporary legal phraseology resulting in an array of different approaches and perspectives.

2 Corpus linguistics phraseology

Much of the fresh impetus for the study of phraseology is rightly attributed to Corpus Lin-
guistics and the emergence of specialised tools and computerised text resources. The term
corpus linguistics phraseology is now used interchangeably with distributional phraseology or
frequency-driven phraseology to indicate a range of methods employed to study the use, dis-
tribution, structure and/or function of multi-word units.² It includes both corpus-based and

² Cf. Grabowski (2015) for the most up-to-date and comprehensive overview of issues involved in corpus
linguistics phraseology research.
corpus-driven methodologies (Tognini Bonelli 2001: 84–87). The corpus-based approach to multi-word units involves pre-selecting such expressions and then analysing the corpus to determine how they are used. For example, one might be able to find out how frequently selected binomial or trinomial expressions (e.g. *null and void* or *give, devise and bequeath*) are used in a specific legal genre, what is their function in context, whether they also appear in other genres, etc. With corpus data tagged with part-of-speech annotation, it is possible to retrieve all instances of such expressions. Such information could be particularly useful for users of legal texts who are not practitioners and/or who do not have sufficient experience with the generic conventions of the texts. In other words, the corpus-based approach can be used to verify one’s intuitions about appropriacy and typicality of particular expressions.

On the other hand, the corpus-driven approach is more bottom-up and inductive. It is also more inclusive, i.e. it aims to provide a full account of possible sequences of words. Thanks to the growing number of custom-designed computer programmes (e.g. kfNgram, ConcGram, etc.) it is possible to generate various contiguous or non-contiguous recurrent word combinations which defy the traditional predefined linguistic categories. Such linguistic constructs have been variously defined and referred to as grammar patterns, lexical bundles, clusters, n-grams, skipgrams, phrase frames, prefabricated patterns, etc. Such constructs are treated as instances of phraseology since there is now general consensus among corpus linguists to treat phraseology as a general term to “describe the tendency of words, and groups of words, to occur more frequently in some environments than in others” (Hunston 2011: 5). However, Gries (2008: 4) notes that the notion of phraseology has become so widespread that many scholars tend to use it without providing clear definitions thus precluding wider applicability and comparability of their work. He offers six criteria useful in defining phraseology: (1) the nature of the elements involved in a phraseologism; (2) the number of elements involved in a phraseologism; (3) the number of times an expression must be observed before it counts as a phraseologism; (4) the permissible distance between the elements involved in a phraseologism; (5) the degree of lexical and syntactic flexibility of the elements involved; and finally (6) the role that semantic unity and semantic non-compositionality/non-predictability play in the definition (Gries 2008: 4). Thus, Gries advocates a more rigorous and principled view of what linguistic constructs should count as phraseologisms and proposes that phraseologism be defined as

> 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. (Gries 2008: 6)

This definition is worth noting for being explicit with regard to each of the parameters and for extending the range of phenomena regarded as phraseologism. As a consequence, it encourages researchers to define carefully the level at which they examine a potential phraseologism and to decide how many elements a phraseologism is supposed to comprise. It also prompts one to consider many types of multi-word expressions as phraseologisms, especially those computer-generated.

However, the influence of corpus linguistics phraseology on legal phraseology extends beyond technological advances in text processing. Rather, corpus linguistics phraseology has paved the way for new and innovative studies that have begun to reveal the potential for investigating various roles and functions performed by phraseologisms in legal discourse.
3 The applications of corpus-based approaches to phraseology in legal discourse

This section highlights some of the most important applications of employing corpus approaches to analyse phraseology in legal discourse taking place during the past fifteen years. These would not be possible without corpora of legal language (see an overview in Pontrandolfo 2012) becoming a fruitful test bed to investigate linguistic and textual features of this specialised language. Legal genres included in the existing corpora vary from legislative to judicial language, from academic to administrative one, thus allowing researchers to undertake monolingual or multilingual contrastive studies on different text types.

One such fundamental application, with significant theoretical underpinnings, concerns constructing new taxonomies of word combinations. Few attempts have been made so far to define the conceptual borders of legal phraseological patterns. Classifications of legal phraseemes have rarely been undertaken in the literature on legal genres, with few remarkable exceptions: Kjær’s classifications of norm-conditioned legal word combinations based on the correlation between their stability and legal constraints (1990: 28–29)\(^3\) as well the more general and term-based typology proposed in 2007 (509–510)\(^4\). More recently, there have been attempts to use the corpus-driven approach to construct classifications of recurrent word combinations according to their functional properties. These include Goźdź-Roszkowski’s cross-genre classifications of legal lexical bundles (2011: 109–142)\(^5\), Biel’s classification of patterns based on a phraseological continuum in legislation (2014: 178–182)\(^6\), and Kopaczyk’s taxonomy of lexical bundles in the early legal discourse of Scottish burghs (Kopaczyk 2013).

Worth reiterating is that the new corpus-driven approach led to the identification of non-terminological units. These classifications represent a successful attempt to define the nature of legal phraseological units – a challenging task – providing a useful framework of analysis that takes into account both a functional and linguistic perspective on legal patterns. Combined with corpus-linguistic methods, they can reveal interesting views on how patterns weave meanings across languages.

Corpus-based research of legal phraseology can be grouped into five major areas (adapted from Pontrandolfo 2013: 151–166):

1) studies that analyse lexico-syntactic combinations in legal language, with a preference for specialised collocations, based on traditional notions of phraseology;
2) studies that focus on the formulaic nature of legal language in terms of routine formulae.

This line of research has been recently augmented by investigations into uninterrupted sequences of word combinations, i.e. lexical bundles, carried out from a number of perspec-

---

\(^3\) (1) Prefabricated word combinations directly prescribed by law; (2) word combinations only indirectly prescribed by law; (3) word combinations based on implicit quotations from other texts in a genre chain in the legal domain; (4) habitually routine phrases (Kjær1990: 28–29).

\(^4\) (1) multi-word terms; (2) collocations with a term; (3) formulaic expressions and standard phrases (Kjær 2007: 509–510).

\(^5\) (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 (Goźdź-Roszkowski 2011: 109–142).

\(^6\) (1) text-organising patterns; (2) grammatical patterns; (3) term-forming patterns (multi-word terms); (4) term-embedded collocations; (5) lexical collocations (Biel 2014: 178–182).
tives: synchronic (e.g. Goźdź-Roszkowski 2006), standardisation of early legal discourse (Kopaczyk 2013), variation in legal discourse (Goźdź-Roszkowski 2011);
3) lexicographic studies aimed at building specialised legal dictionaries (de Groot 1999) with a special focus on terminology and terminography;
4) corpus-based analysis of phraseology applied to contrastive analysis of legal language (Pontrandolfo 2013, Tabares Plasencia 2014) and/or legal translation (Biel 2014);
5) Studies that focus on the way legal patterns weave an intricate web of semantic meanings by resorting to a wider notion of phraseology (Mazzi 2010, Goźdź-Roszkowski/Pontrandolfo 2013, 2014).

These studies have confirmed the relevance of corpus linguistics in discovering recurrent patterns. The order in which these clusters of studies are presented is not casual: a consistent move from the first ones (1, 2, 3) to the last sets of research (4 and 5) is increasingly evident in the literature on legal phraseology. If the first trends served the useful purpose of confirming the highly formulaic nature of legal language, the second group (4 and 5) made a step further. Studying phraseology proved to be extremely important for legal translation purposes: while translating a legal text, phraseology is possibly the element that more than any other (syntax, terminology, etc.) gives the text its ‘legal flavour’, by fulfilling the expectations of the intended readers. It also proved to be crucial in the interpretation of the conceptual relations that can be established between two or more legal terminological units (Meyer/Mackintosh 1994: 339–348) and gave valuable insights not only into legal language but also and most importantly into legal cultures – even more so in the comparison between different legal systems, such as common-law and civil-law ones (cf. Goźdź-Roszkowski/Pontrandolfo 2013).

The classification of studies proposed in this section cannot be regarded as a ‘sealed off’ box with fully defined borders. As a matter of fact, the papers collected in this special issue confirms the enriching interaction and overlaps between the five areas.

4 Synopsis of articles in this issue

The aim of this special issue, made of four original papers, is to collect some original corpus-based research in legal phraseology thus contributing to fill the existing gap in the literature, especially with regard to multilingual studies (both contrastive and translational). The articles collected in this issue bring some further support to the broad conceptualisation of phraseology in LSP studies and highlight new ways in which phraseologisms can be investigated.

The first article “Phraseological profiles of legislative genres: complex prepositions as a special case of legal phrasemes in EU law and national law” is by Łucja Biel who focuses on one of the most typical word combinations in legislative language: complex prepositions. It lies in between the first group of studies (see section 3: analysis of lexico-syntactic combinations in legal language), as it focuses on complex prepositions, a ‘typical’ and widely recognised type of phraseologism, and the forth one (contrastive analysis of legal language and legal translation), as it proposes an interesting interlingual comparison between translated vs. non-translated English and Polish.

The importance of studying complex prepositions relies on the role they play in terms of generic integrity (Bhatia 2004: 115); as a matter of fact, complex prepositions significantly contribute to phraseological profiles of legislative genres. The author analyses the distribution and functions of complex prepositions in multilingual EU law and national law, on a compar-
ative (cross-systemic) and contrastive English-Polish basis, against the background of general language. The analysis relies on three specialised corpora (JRC Acquis for the EU legislation, BoLC for the national legislation of the UK, and PLC for the national legislation of Poland) and two reference corpora (BNC and NKJP). The results confirm that complex prepositions are cognitively salient in the genre of legislation: they show increased distribution against general language, in particular in Polish. It is also demonstrated that national legislation and EU legislation (translated language) are profiled by different sets of salient prepositions, which may adversely affect the readability of the latter due to interference. Functionally, the author demonstrates that the phraseological profiles of legislative instruments are marked by complex prepositions used predominantly in referencing patterns (authority, conflict), conditionals, anchoring (framing) patterns, defining patterns and time deixis.

The second paper “Phraseological units in English-Spanish legal dictionaries: a comparative study” is by Miriam Buendía Castro & Pamela Faber. It belongs to the third group identified in section 3 (lexicographic studies aimed at building specialised legal dictionaries) although it takes into account the translation perspective (group 4).

Based on a translation-oriented lexicographic perspective, the authors analyse the headword *vista* [‘trial, hearing’] in four of the most representative legal English-Spanish dictionaries with a view to evaluating their potential usefulness for translators. In addition to describing the main features of the micro- and macrostructure of the four resources, the authors study how each one deals with the types of phraseological units encoded, the kinds of phraseological information offered, as well as the location of phraseological units within the dictionary. The positive and negative aspects of each resource are highlighted with a view to designing a terminological entry for legal translators that combines the good points of these dictionaries and avoids their drawbacks. The study confirms that a legal dictionary for translators should provide various ways of accessing phraseological units as well as a classification of phraseological information within each entry for a more effective retrieval of information. The authors advocate the inclusion of a short description of the dictionary’s unit so that users are better able to understand its meaning and usage in different contexts as well as its potential contextualised correspondences in the target language and culture.

The third paper “Investigating ‘concgrams’ in the language of contracts and legal agreements” is by Katja Dobrić Basanezé who analyses phraseological units in the language of contracts. The article can be included in the fifth group identified in section 3 (studies that focus on the way legal patterns weave an intricate web of semantic meanings by resorting to a wider notion of phraseology) as it proposes an original corpus-driven investigation of phraseological units in the language of contracts which confirms the formulaic nature of legal language (see group 2 identified in section 3).

The author compares the concgrams found in a bilingual (English, Croatian) comparable corpus of contracts and legal agreements with those included in legal dictionaries and reference corpora. The analysis of the concgrams clustered around keyterms (agreement, parties, consideration, contract and their Croatian counterparts) – i.e. complex nominal phrases, extended lexical collocations, extended prepositional phrases and extended bi- and trinominals – reveal that extended binomials are among the most frequent in the English corpus whereas complex nominal phrases prevail in the Croatian corpus. The results are also seen against the backdrop of the translation solutions provided by legal dictionaries.

The fourth and last paper is “(Re)producing habits in international negotiations: A study on the translation of collocations” by Esther Monzó who explores the subservient habitus
hypothesis explored in Interpreting and Translation Studies, where interpreters and translators are said to be especially sensitive and keen to reproduce social and textual norms, usually dictated by other agents in the field. The paper is an interesting combination of groups 1, 2 and 4 identified in section 3 for the following reasons: it stems from the analysis of a 'traditional' phraseological unit (collocations); it confirms the formulaic nature of legal language; it is mainly translation-oriented/driven as it investigates the translators' behaviours in legal settings.

The author presents a case study of legal translation carried out at the World Trade Organisation (WTO) whose aim is to discover to what extent Spanish legal translators at the WTO reproduce collocations not included in organisational glossaries (which would therefore be prescribed and prescriptive) and to find where differences can be detected. The author combines a semantic target-oriented approach with a corpus-based perspective and focuses on collocations as a key phraseological unit in legal language. The analysis relies on two corpora of WTO agreements: the Spanish translation of panel reports derived from the (lack of) compliance with the Agreement on Subsidies and Countervailing Measures (C1 subcorpus) and the foundational agreements of the WTO (C2 subcorpus), which are taken as the source of textual normality. The collocations under investigation are the result of a comparison with the institutional glossaries used by translators at the WTO: after discarding the phraseological units contained in those resources, the author compares the combinations found in the two corpora. Overall results seem to suggest that C1 is actually taken as a normative source even when there would be room for introducing other options, thus suggesting a move towards standardisation where institutional norms established in authoritative texts (C1) are reproduced in other genres.

5 Looking ahead: future directions

Given the relative scarcity of the multilingual perspective, one should expect to see an increasing presence of translational and contrastive/comparative perspectives in phraseological studies. The growing impact of corpus methodologies for text processing in the legal domain has enabled researchers to generate and analyse a multitude of new types of frequent and recurrent sequences in order to investigate legal discourse from new perspectives. One possible extension of this line of research is to focus on 'units of meaning' rather than sequences of words based on form. This approach is heralded by the concept of a semantic sequence described as a “product of social conditions often requiring similar things to be said, rather than speakers having semi-preconstructed phrases at hand” (Hunston 2008: 292). The emphasis would thus be shifted towards the question of how members of legal professional communities create and understand meanings by means of various phraseological units. In this context, semantic sequences have already begun to be applied to explore the construal of evaluative meanings in judicial discourse (Goźdź-Roszkowski/Pontrandolfo 2013).

Cross-fertilisation between disciplines of empirically-oriented linguistics and LSP studies will result in triangulating the ways in which patterns of co-occurrence are analysed. The coming years should witness an increased integration of research methods, e.g. corpus-based approach combined with Critical Discourse Analysis (CDA) (e.g. Potts/Kjær 2015) thus opening up new avenues for research. Finally, the developments within the area of Digital Humanities which aim at enriching data with more nuanced linguistic and non-linguistic information can be expected to lead to a new generation of corpus data and new applications for them.
References


---

**Prof. nadzw. Dr. hab. Stanisław Goźdz-Roszkowski**
University of Lodz
Department of Translation Studies, Institute of English Studies
Pomorska 171/173, 91-404 Lodz, Poland
gozdz.roszkowski@gmail.com

**Dr. Gianluca Pontrandolfo**
University of Trieste, IUSLIT
Dept. of Legal Science, Language, Interpreting and Translation Studies
Via Filzi 14, 34132 Trieste, Italy
gpontrandolfo@units.it