International Journal of
Language & Law

Hamann/Vogel, JLL 6 (2017): 101–109
www.languageandlaw.de

The Fabric of Language and Law
— Towards an International Research Network for Computer Assisted Legal Linguistics (CAL²)

*Hanjo Hamann and Friedemann Vogel**

Abstract
Law and language can be described as complex institutions with emergent properties, like intricate fabrics woven from single-colored fibers. This metaphor suggests to think of legal language in terms of "patterns": Recurrent motifs in the fabric that the individual language user may not (and in most cases cannot) be aware of, though they explain the development of language more coherently than any narrative based on *a priori* rules. This perspective corresponds with the recent trend towards computer linguistics using "text as data". To discuss how these approaches might impact research on the language of law, the Heidelberg Academy of Sciences and Humanities hosted the first international conference on "The Fabric of Language and Law" from the perspective of legal corpus linguistics. Selected papers presented at this meeting in March 2016 were subsequently peer-reviewed and published in an eponymous volume of the International Journal of Language & Law (JLL), edited by the present authors as convenors of the conference. This special issue introduction elaborates on the topic of this meeting, summarizes its contributions, and contextualises the publications that resulted from it. The authors hope that this exchange, which has meanwhile been continued across the Atlantic, may help to establish an international network for research on Computer Assisted Legal Linguistics (CAL²).

Keywords
corpus, computer linguistics, semantics, law and language, legal linguistics, big data, CAL²

Editorial (not reviewed), published online: 7 September 2017

* *Hamann*: Max Planck Institute for Research on Collective Goods, Bonn (Germany), hamann@coll.mpg.de; *Vogel*: University of Freiburg (Germany), friedemann.vogel@mkw.uni-freiburg.de. The authors thank Yinchun Bai and Isabelle Gauer for their assistance, and the Heidelberg Academy of Sciences and Humanities for generous financial support through their WIN funding programme.

1. Legal Language as a Fabric

"What we call chaos is just patterns we haven't recognized.
What we call random is just patterns we can't decipher."
(Chuck Palahniuk, *Survivor* 1999, p. 118)

What do law and language have in common?

To the untrained eye, both may occasionally seem erratic or even chaotic. Think, on the one hand, of the supposed "unique lunacy of the English language" (Lederer, 1990) or, on the other, of ubiquitous collections of "famous wacky laws", which often turn out to be "not-so-wacky" at all (McClurg, 2011). There may be a deeper reason for language and law being likewise accused of feeblemindedness: Both can be described as "phenomena of the third kind" (Keller, 1990) – not growing entirely rank (as autonomous organisms would do), but not constructed to plan either (as artifacts would be). This was previously emphasized by Hamann (2017: 181) who argued that both law and language are emergent systems – emerging from theory-based rules not by way of arithmetic or logic, but by collective habits producing *patterns* of usage. Considering further that law can only be conceived of through and in language,[1] it even forms a second-order usage pattern: Law is one manifestation of how we use rules and norms stated in the form of language, itself being our way of using semantic symbols and signs. A fitting metaphor might be that of a cross-woven *fabric*.

The English word *fabric* – meaning a "thing made; a structure of any kind" – dates from the late 15th century, but came down to us all the way from a Proto-Indo-European word for "fitting together" or "fashioning", via its 1791 usage for "textile, woven or felted cloth" (etymonline.com, 27 Aug 2017). If we think about *legal language as a fabric* then, we don't just emphasize its human-made aspects, we also suggest more specific similarities between the way textiles are fashioned and the way legal language is. Think about a texture made by interweaving fibers: The woven cloth cannot exist without a self-stabilizing structure of single fibers. It is a skeleton: Fibers do not stick together; they keep hold of each other and equilibrate *as a part* of the fabric. The whole exists only as an interaction of its fibers.

Language, too, is a fabric: We do not use a word (or phrase, or text) in isolation, but always grounded in a specific communication setting (Barsalou, 2008; Clark & Brennan, 1993; Glaser, Strauss & Paul, 1967/2008): *who* (speaker alone or together with others), *when* (current day as well as historical period), *where* (formal versus familiar; cultural location), *to whom* (addressees and recipients), *through which medium* (face to face or via e-mail, chat, etc.), and so on. Besides, any expression of language is located in a stream of other expressions, connected with earlier and subsequent words, phrases, paragraphs, etc. Each expression can only exist and be "meaningful" in relation to the

[1] Not necessarily through and in *texts*, as Thilo Kuntz helpfully pointed out citing Sachs (forthcoming 2017).

given circumstances in time and space. In other words: The usage of a particular word is an intentional selection of alternatives, selected according to its co-text as well as its social context (see Gumperz, 1982; Wittgenstein, 2003 [1953]). A single word is like a single fiber, while the whole communication setting, the entire text including producer and audience, constitutes a fabric. On the second level, law also constitutes a hypertextual network of references between the world of legal norms, the world of everyday life and the world of texts (Vogel, Hamann & Gauer 2017: fig. 1). In other words: Law is text, law is intertextuality (Morlok, 2004; 2015).

For neither of the two layers of fabric does its weaver see the entire canvas, as German poet Heinrich Heine described so beautifully in the mid-19th century (see Hamann & Vogel, 2017: 87, referring to Heine 1851/1905: 18):

> "Years, revolving, come and vanish;
> To and fro the spool is humming
> In the loom, and never resting;
> What it weaves no weaver knows."

2. Legal Language as Big Data

If different communication settings produce different fabrics, does this mean that any fabric is unique?

Not entirely. Our cognitive capacities simply do not permit to parse every utterance bit by bit, word by word – we would never be able to communicate. We do, however, communicate successfully because our language is full of *patterns*: multi-word-units with idiomatic notions (Steyer, 2013), speech stereotypes (Feilke, 1989), speech sequences or procedures for turn-taking to manage discourses effectively (Goffman, 1983; Sacks, Schegloff & Jefferson, 1974). So once we behold our fabric at medium range, we can observe its regularly recurrent patterns. Yet these patterns are not properties of the fabric itself, but result from the patterns of our perception and cognition, such as frames and scripts (see Barsalou, 1992; Minsky, 1975; Schank & Abelson, 1977) or stereotypes and prejudices. Even in science, patterns (*prototypes, clusters, sort/kind of, genre* and so on) play an important role and are essentially the basis of any hypothesis. In the case of law and language, their re-cognizable patterning enables us to approach them systematically and, therefore, empirically.

These realizations coincide with a fundamental change in the context of language and law over the last twenty years: The digitalization of all areas of life changed the production of legal fabric as well as our practices of language patterning (see Vogel, 2015). Their fiber structure becomes more easily cognizable and even explorable: Intertextuality, references, etc. are now "clickable" through hypertext and hypermedia. More and more legal texts are saved in digital databases, available through search en-

gines, and judges use software to manage formulaic text modules for their decisions. This digital trend also proffers new potential for legal linguistics: It may turn to computer-assisted methods, as text has become data.[2]

Computer supported corpus linguistics has developed all over the world for the past 30 years (see McEnery & Wilson, 1997; Teubert, 2004; Lüdeling & Kytö, 2008). Corpus linguists use algorithms and software developed by computational linguistics and computer scientists to statistically discern language patterns at various levels. Epistemologically, two approaches may be used: *Corpus-based* approaches usually seek to test qualitative hypotheses, for example, using frequency analysis of an expression in selected text collections. In contrast, *corpus-driven* approaches try to let the corpus speak for itself (see Tognini-Bonelli, 2001), so researchers calculate various parameters and try to develop new hypotheses grounded in the corpus. Both approaches are extremes on a gradated spectrum, i.e., most corpus linguists use both corpus-based and corpus-driven methods (Fillmore, 1992; Stefanowitsch, 2008).

The decisive advantage of these computer supported methods is to control intuition. Though native speakers' intuition is an irreplaceable presupposition for qualitative assumptions about language use, intuition sometimes fails or is at least not adequate – especially for estimating the frequency of phenomena. In such cases computers are simply better. On the other hand algorithms cannot understand semantic structures of the data they analyze, so they cannot supplant qualitative reasoning. In this sense, one of the most recent approaches came to be labelled "computer *assisted* legal linguistics" (Vogel, Hamann & Gauer 2017; Hamann & Vogel, forthcoming 2017).

3. Fabric of Language and Law – The Conference

These themes inspired a conference in March 2016, being the first international meeting to bridge corpus linguistics and law. Hosted by the Heidelberg Academy of Sciences and Humanities' research group "Computer Assisted Legal Linguistics" (CAL²), it was entitled *The Fabric of Language and Law. Discovering Patterns through Legal Corpus Linguistics* and drew an audience of some forty participants to Heidelberg (Germany).

Speakers and participants from Germany, Switzerland, Italy, Poland, Spain and the U.S. (most from language sciences, law, philosophy and computer science) gathered for two days, attending a total of ten invited talks and a concluding panel discussion. Speakers included Larry Solan, Stephen Mouritsen, Łucja Biel, Stanisław Goźdź-

[2] This (possibly overused) trope may be substantiated by casually observing that the Department of Politics at Princeton University has hosted eight "Text as Data Conferences" (q-aps.princeton.edu/news/text-data-conference), the College of Social Sciences and Humanities at Northeastern University hosted seven "New Directions in Analyzing Text as Data" conferences (northeastern.edu/textasdata2016), and academic papers from various disciplines all use "text as data" in their title.

Roszkowski, Stefan Höfler, Ruth Breeze, María José Marín, Giulia Venturi, Rema Rossini Favretti and the conference's convenors Hanjo Hamann and Friedemann Vogel. On the final panel, Solan and Biel were joined by Dieter Stein and Andreas Abegg. A more detailed summary of the conference schedule was previously reported by Vogel et al. (2016), an article-length conference report by Lukas (2017).

Following the conference, its speakers were invited to submit full-length papers which were then peer reviewed for publication in JLL. This resulted in five JLL publications in its 2017 "Fabric of Language and Law" volume, which are summarized and contextualized in the following section. The debate has meanwhile continued on the other side of the Atlantic, with two of the Heidelberg contributors, as well as one of the present authors, joining a variegated roster of U.S. scholars for the second international conference on law and corpus linguistics, hosted by Brigham Young University in Provo, Utah – see the pending 2017 special issue of *BYU Law Review*.

4. Taking Stock of Legal Linguistics

In his keynote paper entitled *Patterns in Language and Law*, law professor and U.S. legal linguistics pioneer Larry Solan (2017b) builds on Pinker's (1999) distinction between rule-like and pattern-like structures of language and shows that law can be conceptualized in similar terms. As one of the most prolific advocates for legal linguistics, Solan is also one of the first to extensively incorporate corpus methods into his research (see Solan, 2016; Solan & Gales, 2016; Solan, 2017a; Solan & Gales, forthcoming 2017). He shows how the concepts of corpus linguistics may help to clarify and rethink four perennial problems of legal theory: The "inevitability of standards within rules"; coherence reasoning as "a basic rule of law value"; the kinship between ordinary meaning inquiry and "category membership and goodness of fit"; as well as "laws that explicitly call for pattern-like interpretation". Using U.S. court cases as illustrations, the author also reveals how patterns affect legal language and adjudication. From this analysis, he concludes that "corpus analysis cannot solve all of the legal system's interpretive puzzles" but reveals the surprising and not yet fully theorized extent to which "statutory analysis in law is based on the notions of central tendency and goal orientation".

These theoretical macro-reflections are then contextualized in another paper, by U.S. legal corpus linguistics pioneer Stephen Mouritsen (2017). In his paper on *Corpus Linguistics in Legal Interpretation* as *An Evolving Interpretative Framework*, he analyses and documents the development of the field within the U.S. and provides the much-needed origins narrative that the field had yet been missing (see Hamann & Vogel, forthcoming 2017). The author may be the best-placed of all people to relate this story, as it was his own work which inspired the movement (Mouritsen, 2010; 2011) at around the same time that German legal scholars started using corpus analysis (Kudlich & Chris-

tensen, 2009) and legal linguists started developing a coherent interdisciplinary methodology (Felder, Müller & Vogel, 2010; Vogel, 2012a; 2012b). In the U.S., according to Mouritsen, legal corpus linguistics ("LCL") started with judges succumbing to their "data impulse": By using "quasi-corpora", they inspired an actual wave of corpus usage in statutory interpretation, which eventually even made it into legal training at one U.S. law faculty. The article concludes with an extensive discussion of potential challenges to the use of corpora in law, showing how much reflection remains yet to be done (see also Lee & Mouritsen, forthcoming 2017).

Building on this theoretical groundwork, Spanish linguist María José Marín (2017) takes a more hands-on approach towards *Legalese as Seen Through the Lens of Corpus Linguistics*. Her thorough review of computer linguistic methodology as well as extensive software tests informed the author's *Introduction to Software Tools for Terminological Analysis*. Comparing various algorithms for automatic term recognition ("ATR"), the author provides an instructive and quite rich summary of the technological state of the art. Her text is illustrated with examples from the author's own "British Law Report Corpus" (BLaRC) which had already been introduced in previous studies (Marín & Rea Rizzo, 2012; Marín, 2014). This corpus-driven application makes the text easily accessible even to the computer linguistic novice, and hints at a wide array of applications that will further expand as more research is carried out and improved software tools become available, as the author notes in concluding.

One of the most important next steps for corpus linguistics in law is then paved by British philologist Ruth Breeze (2017) in her study on *Corpora and Computation in Teaching Law and Language*. Extending previous work by the same author (Breeze, 2015) and others (Hafner & Candlin, 2007), she shows how corpora can be used to facilitate language acquisition and terminology training in a particularly important legal domain: Business law. If law students become familiar with the concepts and methods of corpus research at an early stage of their education, this will not only change their concept of legal language ("application" of language "laws" vs. inductive analysis of usage patterns), but also enrich their methodological toolbox in quite tangible ways. In this sense, new teaching methods for students of law and language may be key to the dissemination and acceptance of the new methodology. This insight ties her contribution to Mouritsen's (2017), who had introduced corpus methods into his law school's curriculum, thus reaffirming the demand perceived by Breeze.

To round off the conference's special issue, JLL republishes a transcript of the final panel discussion that was previously published in Vogel et al. (2016). Dieter Stein, as a founding member of the International Language and Law Association (ILLA), chaired an open discussion involving conference speakers Solan (also ILLA co-founder) and Biel, joined on the podium by Swiss legal theorist Andreas Abegg. They were asked to first summarize their "lessons learned" at Heidelberg, and then discussed the present state of the art in corpus linguistics with the audience. One of the audience members, in citing "Alice in Wonderland", unwittingly coined the panel discussion's published ti-

tle: *"Begin at the beginning". Lawyers and Linguists Together in Wonderland.* Its transcript both documents the conference's bottom line and inspires future debate on essential epistemological issues of interdisciplinary research on law and language, and evidence-based policy (see Hamann & Vogel, forthcoming 2017).

References

Barsalou, Lawrence W. (1992). Frames, Concepts, and Conceptual Fields. In Lehrer & Kittay (Eds.), *Frames, fields, and contrasts. New essays in semantic and lexical organization* (pp. 21–74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Barsalou, Lawrence W. (2008). Grounded Cognition. *Annual Review of Psychology, 59*, 617–645. DOI: 10.1146/annurev.psych.59.103006.093639.

Breeze, Ruth (2015). Teaching the Vocabulary of Legal Documents: A Corpus-Driven Approach. *Journal of English for Specific Purposes at Tertiary Level (ESP Today), 3*(1), 44–63. Available at esptodayjournal.org/pdf/current_issue/2015/3. RUTH BREEZE - full text.pdf.

Breeze, Ruth (2017). Corpora and Computation in Teaching Law and Language. *International Journal of Language & Law, 6*, 1–17. DOI: 10.14762/jll.2017.001.

Clark, Herbert H. & Brennan, Susan E. (1993). Grounding in communication. In Resnick, Levine & Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Feilke, Helmuth (1989). Funktionen verbaler Stereotype für die alltagssprachliche Wissensorganisation. In Knobloch (Ed.), *Kognition und Kommunikation. Beiträge zur Psychologie der Zeichenverwendung* (pp. 71–84). Münster: Nodus.

Felder, Ekkehard, Müller, Marcus & Vogel, Friedemann (2010). Das Heidelberger Korpus – Gesellschaftliche Konflikte im Spiegel der Sprache. *Zeitschrift für Germanistische Linguistik (ZGL), 38*, 314–319.

Fillmore, Charles J. (1992). 'Corpus linguistics' vs. 'Computer-aided armchair linguistics'. In Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (pp. 35–60). Berlin/Boston: Mouton de Gruyter.

Glaser, Barney G., Strauss, Anselm L. & Paul, Axel T. (1967/2008). Grounded Theory: Strategien qualitativer Forschung (German 2nd ed. 2008). Bern: Huber.

Goffman, Erving (1983). The Interaction Order. *American Sociological Review, 48*(1), 1–17.

Gumperz, John J. (1982). *Discourse Strategies.* Cambridge, UK: Cambridge University Press.

Hafner, Christoph A. & Candlin, Christopher N. (2007). Corpus Tools as an Affordance to Learning in Professional Legal Education. *Journal of English for Academic Purposes, 6*(4), 303–318. DOI: 10.1016/j.jeap.2007.09.005.

Hamann, Hanjo (2017). Strukturierende Rechtslehre als juristische Sprachtheorie. In Felder & Vogel (Eds.), *Handbuch Sprache im Recht* (pp. 175–186). Berlin: de Gruyter. DOI: 10.1515/9783110296198-009.

Hamann, Hanjo & Vogel, Friedemann (2017). Die kritische Masse. Aspekte einer quantitativ orientierten Hermeneutik am Beispiel der computergestützten Rechtslinguistik. In Schweiker et al. (Ed.), Messen und Verstehen in der Wissenschaft. Interdisziplinäre Ansätze (pp. 81–95). Wiesbaden: J.B. Metzler (Springer imprint). DOI: 10.1007/978-3-658-18354-7_7.

Hamann, Hanjo & Vogel, Friedemann (forthcoming 2017). Evidence-Based Jurisprudence meets Legal Linguistics. Unlikely Blends Made in Germany. *Brigham Young University Law Review, 43*.

Heine, Heinrich (1851/1905). Hebrew Melodies. Third Book, Jehuda Ben Halevy (Fragment). Translated by Armour (Ed.), The Works of Heinrich Heine XII: Romancero Book III. London: William Heinemann. Available at archive.org/details/worksofheinrichh12hein.

Keller, Rudi (1990): *Sprachwandel. Von der unsichtbaren Hand in der Sprache.* Tübingen: Francke. [English version: On Language Change. The Invisible Hand in Language 1994; German 4th ed. 2014].

Kudlich, Hans & Christensen, Ralph (2009). *Die Methodik des BGH in Strafsachen*. Köln: Heymanns.

Lederer, Richard (1990). Crazy English: the Ultimate Joy Ride Through Our Language. New York: Pocket Books. Cited from excerpt at academic.luzerne.edu/shousenick/101--EXAMPLE_EnglishCrazyLanguage_Lederer.doc.

Lee, Thomas R. & Mouritsen, Stephen C. (forthcoming 2017). Judging Ordinary Meaning. *Yale Law Journal, 126.*

Lüdeling, Anke & Kytö, Merja (Eds.) (2008). *Corpus Linguistics: An International Handbook.* Berlin: de Gruyter.

Lukas, Christoph (2017). Korpuslinguistik und Recht. Bericht über die Konferenz „The Fabric of Law and Language" der Heidelberger Akademie der Wissenschaften vom 18. und 19. März 2016. *Archiv für Rechts- und Sozialphilosophie, 103*(1), 138–145. Available at ingentaconnect.com/contentone/fsv/arsp/2017/00000103/00000001/art00007.

Marín, María José (2014). A Proposal to Exploit Legal Term Repertoires Extracted Automatically from a Legal English Corpus. *Miscelánea: A Journal of English and American Studies 49*, 53–72. Available at miscelaneajournal.net/index.php/misc/article/view/177.

Marín, María José (2017). Legalese as Seen Through the Lens of Corpus Linguistics. An Introduction to Software Tools for Terminological Analysis. *International Journal of Language & Law, 6*, 18–45. DOI: 10.14762/jll.2017.018.

Marín, María José & Rea Rizzo, Camino (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. Available at publicaciones.unirioja.es/ojs/index.php/jes/article/view/184/164.

McClurg, Andrew Jay (2011). Famous Wacky Law Exposed as Not-So-Wacky. *McClurg's Legal Humor: Legal Mythbusters*, 25 Nov. Retrieved from lawhaha.com/famous.

McEnery, Tony & Wilson, Andrew (1997). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Minsky, Marvin (1975). A framework for representing knowledge. In Winston & Horn (Eds.), *The psychology of computer vision* (pp. 211–277). New York: McGraw-Hill.

Morlok, Martin (2004). Der Text hinter dem Text. Intertextualität im Recht. In Blankenagel, Pernice & Kotzur (Eds.), *Verfassung im Diskurs der Welt. Liber Amicorum für Peter Häberle zum siebzigsten Geburtstag* (pp. 93–136). Tübingen: Mohr Siebeck.

Morlok, Martin (2015). Intertextualität und Hypertextualität im Recht. In Vogel (Ed.), *Zugänge zur Rechtssemantik* (pp. 69–90). Berlin: de Gruyter.

Mouritsen, Stephen C. (2010). The Dictionary Is Not a Fortress: Definitional fallacies and a corpus-based approach to plain meaning. *Brigham Young University Law Review, 2010*, 1915–1978. Available at digitalcommons.law.byu.edu/lawreview/vol2010/iss5/10.

Mouritsen, Stephen C. (2011). Hard Cases and Hard Data: Assessing corpus linguistics as an empirical path to plain meaning. *Columbia Science and Technology Law Review, 13*, 156–205. Available at stlr.org/cite.cgi?volume=13&article=4.

Mouritsen, Stephen C. (2017). Corpus Linguistics in Legal Interpretation. An Evolving Interpretative Framework. *International Journal of Language & Law, 6*, 67–89. DOI: 10.14762/jll.2017.067.

Pinker, Steven (1999). *Words and Rules: The Ingredients of Language*. New York: Basic Books.

Sachs, Stephen E. (forthcoming 2017). Originalism Without Text. *Yale Law Journal, 127.*

Sacks, Harvey, Schegloff, Emanuel A. & Jefferson, Gail (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language, 50*(4), 696–735. DOI: 10.2307/412243.

Schank, Roger C. & Abelson, Robert P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Solan, Lawrence M. (2016). Can Corpus Linguistics Help Make Originalism Scientific? *Yale Law Journal Forum, 126*, 57–64. Available at yalelawjournal.org/forum/can-corpus-linguistics-help-make-originalism-scientific.

Solan, Lawrence M. (2017a), Linguistic knowledge and legal interpretation – what goes right, what goes wrong. In Slocum (Ed.), *The Nature of Legal Interpretation: What Jurists Can Learn about Legal Interpretation from Linguistics and Philosophy* (pp. 66–87). Chicago: University of Chicago Press.

Solan, Lawrence M. (2017b). Patterns in Language and Law. *International Journal of Language & Law, 6*, 46–66. DOI: 10.14762/jll.2017.046.

Solan, Lawrence M. & Gales, Tammy (2016). Finding ordinary meaning in law: The judge, the dictionary or the corpus. *International Journal of Legal Discourse, 1(2)*, 253–276. DOI: 10.1515/ijld-2016-0016.

Solan, Lawrence M. & Gales, Tammy (forthcoming 2017). Corpus Linguistics as a Tool in Legal Interpretation. *Brigham Young University Law Review, 43*.

Stefanowitsch, Anatol (2008). Konstruktionsgrammatik und Korpuslinguistik. In Fischer, Stefanowitsch & Fischer-Stefanowitsch (Eds.), *Konstruktionsgrammatik I. Von der Anwendung zur Theorie* (2nd ed.) (pp. 151–175). Tübingen: Stauffenburg.

Steyer, Kathrin (2013). *Usuelle Wortverbindungen: Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

Teubert, Wolfgang (2004). Language and corpus linguistics. In Halliday (Ed.), *Lexicology and corpus linguistics. An introduction* (pp. 73–112). London: Continuum Books.

Tognini-Bonelli, Elena (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: J. Benjamins.

Vogel, Friedemann (2012a). Das Recht im Text: Rechtssprachlicher Usus in korpuslinguistischer Perspektive. In Felder, Müller & Vogel (Eds.), *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen* (pp. 314–353). Berlin: de Gruyter.

Vogel, Friedemann (2012b). *Linguistik rechtlicher Normgenese. Theorie der Rechtsnormdiskursivität am Beispiel der Online-Durchsuchung*. Berlin: de Gruyter.

Vogel, Friedemann (Ed.) (2015). *Zugänge zur Rechtssemantik: Interdisziplinäre Ansätze im Zeitalter der Mediatisierung zwischen Introspektion und Automaten*. Berlin: de Gruyter.

Vogel, Friedemann, Hamann, Hanjo & Gauer, Isabelle (2017). Computer-Assisted Legal Linguistics. Corpus Analysis as a New Tool for Legal Studies, *Law & Social Inquiry, 42*, early view. DOI: 10.1111/lsi.12305.

Vogel, Friedemann; Hamann, Hanjo; Stein, Dieter; Abegg, Andreas; Biel, Łucja & Solan, Lawrence (2016). Begin at the Beginning: Lawyers and Linguists Together in Wonderland. *The Winnower, 3*, 4919. DOI: 10.15200/winn.148184.43176.

International Journal of
Language & Law

Solan, JLL 6 (2017): 46–66
www.languageandlaw.de

Patterns in Language and Law

Lawrence M. Solan[*]

Abstract
Our language faculty is rule-like in some ways, pattern-like in others, as Steven Pinker (1999) has shown. Much of syntax is describable a set of rules, whereas the range of meanings attributed to a word is best described in terms of patterns. Laws are typically written as rules, but they are written in words, many of which display pattern-like arrays of usage. Legal systems default to an expression's "ordinary meaning," requiring estimates of patterns of usage. Recently, advances in corpus linguistics have been adduced by judges and legal scholars in this regard. Furthermore, open-textured legal terms, including the word "pattern" itself, are by their nature more describable in terms of patterns of their application than in terms of hard-and-fast rules. Apart from linguistic issues in legal interpretation, legal systems value coherence, requiring that like things be treated alike, often focusing on patterns of how laws are applied. At times, however, these patterns uncover biases in a law's application. This article attempts to describe how this duality in both linguistic description law interact with each other.

Keywords
corpus, ordinary meaning, rule, pattern, standard, legal linguistics, courts, US judiciary

Submitted: 29 October 2016, accepted: 25 July 2017, published online: 24 August 2017

[*] Brooklyn Law School, USA, larry.solan@brooklaw.edu. The author expresses his gratitude to the editors and two anonymous reviewers for their helpful comments on an earlier draft of this article.

1. Introduction

The age of big data has not only found its way to the study of language and to the study of law, but it has also found its way to the interdisciplinary field of legal linguistics. The use of linguistic corpora in legal analysis is growing, both in the determination of individual cases and in the study of language use that reveals regularities that are not part of the "official" canon of legal doctrine (see, e.g., Vogel, Hamann & Gauer, 2017).

This article aims to delineate the aspects of legal reasoning and our linguistic competence that combine to make this development possible. The most interesting cases come when the law – which is most often organized as a rule-like system – responds to patterns of family resemblance, rather than to absolute criteria. Legal theorists from both common law and Roman law traditions embrace coherence as a basic rule of law value. Because it is not possible to articulate precisely what common features produce the kind of coherence the law values, the effort to achieve coherence produces legal doctrines that are always subject to challenge.

The law also resorts to pattern-like interpretation when it defaults to the "ordinary meaning" of words and phrases in legislation. Most significant for our purposes, "ordinary meaning" is a distributional fact. A meaning is "ordinary" if it occurs commonly (how commonly is a matter of some disagreement). There is no better way to determine how commonly a particular meaning is assigned to a word than to review a large corpus of general usage of the language and to compute the relative frequency of occurrence. Indeed, excellent work in this regard by Stephen Mouritsen (2010, 2011), and more recently Lee & Mouritsen (forthcoming) demonstrates the utility of corpus analysis in determining ordinary meaning in legal cases, replacing the intuitions of judges and reference to dictionaries as methods in legal decision making that depends upon ordinary meaning. In addition, several U.S. judges have actually employed the method in their resolution of legal disputes.

Moreover, while linguists have successfully made great progress over the past half century using their intuitions as native speakers as the data on which to create theories that separate the grammatical from the ungrammatical, it is much less clear that linguists, or others for that matter, have good intuitions about distributional facts. On the contrary, people are subject to "false consensus bias" when it comes to the distribution of meanings. As native speakers, we tend to think that our understandings of words are the normal ones, not recognizing the possibility that we are outliers. In a study conducted with both lay people and sitting judges, Solan, Rosenblatt & Osherson (2008) asked participants whether they considered employees who were sickened by inhaling sand particles in a factory that used sand to make its equipment to have suffered a "pollution injury," which is a question relevant to in insurance law. About 45 percent of lay people said it was pollution, about 42 percent said it was not, and about 13 percent said they could not decide. All three groups, when told that we had presented the same scenario to 100 people just like them overestimated the percentage of peo-

ple who agreed with their judgment. The first two groups estimated that they were in agreement with 60–65 percent of respondents; those who said they could not decide estimated 38 percent agreement. Judges were also subject to this false consensus bias. While only a small percentage (12 percent) of judges thought the injury was caused by pollution, those that so responded estimated that about 70 percent of judges would agree with them.

A final kind of distributional fact plays a significant role in the relationship between language and law. Laws are not always applied even-handedly, without regard to such things as race and gender. For example, DWI (driving while intoxicated) is an infraction everywhere in the U.S. DWB (driving while black) is not supposed to be an infraction at all. Yet the expression developed to describe differential treatment of racial and ethnic groups by the police. Some instances of disparate treatment are linguistic in nature. When people have the option of choosing among different words or different linguistic structures to convey a thought, the choices they make may reveal various schemas that they carry in their minds (see Shuy, 2015). The distribution of those choices may further reveal sociologically-interesting generalizations. Some relate to hot-button issues, such as the appropriate vocabulary to use in the realm of immigration and diversity. I return to such cases at the end of this article.

2. Rules and Patterns in Language

The human language faculty is multidimensional. Part of it is a rule-like system. Take, for example, the classic case of regular plural formation in English.

book	books
leg	legs
glaze	glazes

While the spelling is the same across these examples, whether the sound of the plural is [s], [z] or [Iz] depends upon the final sound of the word in its singular form. Big data will not be very useful here, other than to reveal potential dialect differences with respect to a small set of words that some may regard as regular, others as irregular. If someone were to say "bookes" as the plural of book, we would conclude that the person is not competent in applying the English plural rule. For that reason, it would not contribute to knowledge about the structure of English if it is discovered from a corpus that someone has made such a mistake in writing. In other words, regular plural formation in English does not produce a set of distributional facts.

The seminal work on this dual nature of linguistic phenomena is Steven Pinker's (1999) book, *Words and Rules*. Pinker argues that some aspects of language – irregular forms, vocabulary and word meaning in particular – must be learned individually,

while other aspects are rule-governed. Pinker distinguishes between the regular plural forms and the irregular ones (*fish*, *woman*, *man*, etc.) that children must learn one by one. German plural forms are more word-like in nature because there are numerous classes of plural forms. Which class a word belongs to is largely (but not entirely) unpredictable from its sound, although as Yang (2002), points out, there may be more regularity in this regard than is sometimes thought.

Furthermore, some aspects of language are best described as independent constructions (Goldberg & Jackendoff, 2004), or as lexical bundles (Conrad & Biber, 2004) and are not derivable compositionally. These too are subject to distributional analysis. As Jackendoff (2008) puts it, words and rules form a continuum, rather than two distinct discreet sets of linguistic phenomena.

For the most part, syntax is rule-based. Consider "do support" in English, studied routinely in first year linguistic courses that use English as an example language.

(1) *Visited you your mother yesterday?

(2) Did you visit your mother yesterday?

(3) Have you visited your mother recently?

(4) *Did you have visited your mother recently?

Sentences of the form (1) are grammatical in many languages, including German and the romance languages. But they are not grammatical in English: the form (2) is required. Moreover, when we hear questions like (2) we have no difficulty knowing what is being questioned. Both the form and the interpretation are fixed, at least to that extent. In English, we insert *do* when a finite verb is being questioned, but not when there is an auxiliary verb that may invert with the subject. We also insert *do* before negation:

(5) Bill didn't leave.

(6) *Bill not left.

(7) ?Bill left not.

Here again, once a child internalizes the system, there is no need to worry about what the main verb is, because it does not matter. The rule applies across all main verbs.

How do the pattern-like aspects of language emerge? First, our computational rule-based linguistic system often leaves a great deal of flexibility in how words are used, and which words and structures are used. This flexibility licenses usage to be distributed in patterns, both within a single individual and across the population of competent speakers of the language. It further allows for patterns within discreet genres, such as the various ways that language is used in the legal system. Returning to our example, while do-support is both rule-driven and obligatory, whether one asks many or only a few questions, or how frequently one uses negation is not.

Gaining an understanding of the linguistic patterns in legal language provides an excellent opportunity for corpus analysis. For one thing, examining a large body of da-

ta permits the researcher to uncover actual distributional facts beyond intuitions developed from exposure to small samples. As Goźdź-Roszkowski (2011: 34) describes it:

> "Corpus linguistics is rightly viewed as a research approach that has developed over the past 40 years to study *language use* in large, principled collections of texts. The central goal of corpus-based analysis is to document and interpret generalizable patterns of use" (emphasis added).

Of course, people generally use language within the bounds of what their rules of grammar permit. For this reason, it is helpful at this point to look briefly at the architecture of the language system. It is here that the generative grammar approach and corpus linguistic analysis of the patterns of usage can meet productively. Let us assume that the human language faculty has at least the following (taken roughly from Jackendoff, 2003 and Chomsky, 2005) as its design (see Solan, 2017, for further detail): A computational system that generates well-formed structures; A relationship between these structures and meaning; An interface between sound and meaning, mediated by the computational system, so that we can break the flow of speech up into words and phrases and use this information to interpret what we hear; An interface between the computational system and its interpretation on the one hand, and a conceptual system on the other, so that we express in words and phrases the concepts we intend to express; Interfaces with various inferential systems that rapidly place the language we use in sufficient context to make sense of it (discourse, pragmatics, cultural assumptions, etc.).

Virtually all of the patterns in language occur at the interfaces between the computational component of language and other linguistic and non-linguistic cognitive systems, especially with respect to conceptualization and our inferential system. To the extent that words have multiple senses, and to the extent that they are polysemous, it is possible to ascertain central tendencies and to privilege those as the ones most likely intended by users. The same holds true for inferences that are typically drawn from language in particular circumstances. People know what a guest has in mind when she asks, "is there any salt" but we can never be certain that different people draw the same inferences.

As for how we conceptualize, research in cognitive psychology for the past forty years has demonstrated convincingly that people judge certain items as better members of categories. In this sense, category membership is graded. Most of the work in this area privileges the "prototype" as the best example (see, e.g., Rosch, 1975), although some work seems to show that items that best match a category's goal are judged better examples even when they are less typical (Lynch, Coley & Medin, 2000; Barsalou, 1983). This is not to say that category membership is determined by typicality of membership (Armstrong, Gleitman & Gleitman, 1983). People judge penguins and robins as both being birds and do not believe that "birdhood" is a graded category. Yet robins, at least in western culture, are considered better examples of birds than are penguins. Moreover, failure in matching words and concepts occurs when we find ourselves using a word or phrase that does not actually communicate what we are trying to convey, and in instances of vagueness, where what we try to convey is on the borderline between concepts.

3. Rules and Patterns in Law

Just as language is structured as a mixture of rules and patterns, so is the law. Legal systems are actually structured as a series of rules, even more so in code-driven legal systems, such as those of Europe. Yet these rules are comprised of words, and the words have prototype effects and distributional properties that create patters of usage. Thus, much of the pattern-like nature of legal analysis is linguistic in nature. The pattern-like nature of law, though, goes way beyond questions of word meaning. Coherence is a fundamental legal value in its own right, as theorists from many different perspectives have noted (see, e.g., Dworkin, 1986; Shapiro, 2011; Zippelius, 2008). First, we look at law as a set of rules.

3.1. The Rule-Like Nature of the Law and the Inevitability of Standards within Rules

Laws – and penal laws in particular – are generally written as classical definitions. The burglary law defines burglary; the arson law defines arson. The laws are comprised of a list of "elements" all of which must obtain for the law to apply to a given situation. Each element is necessary, and together they are sufficient to define what is proscribed. The elements, in turn, are presented either conjunctively or as part of a list of which at least one member must obtain. Thus, legal rules can be described using only the Boolean operators, "and," "or" and "not."

As Schauer (2009) points out, we generally conceptualize the law as a collection of rules. For example, the common law definition of burglary was breaking and entering into someone else's dwelling at night with the intent to commit a felony therein. Modern statutes expand the crime to include any building and not to require that the crime occur in the nighttime. Thus, stealing tomatoes from the garden of another person is not burglary, although it is otherwise against the law. Breaking into a factory to deface it is now an act of burglary, but was not an act of burglary at common law. And so on.

This rule-like nature of laws is not confined to common law jurisdictions. Consider the perjury laws, Sections 153 and 154 of the German Criminal Code. Section 153 covers unsworn false testimony:

> Whosoever as a witness or expert gives false unsworn testimony before a court or other authority competent to examine witnesses and experts under oath shall be liable to imprisonment from three months to five years.

Section 154 defines perjury as falsely swearing an oath to tell the truth:

> (1) Whosoever falsely takes an oath before a court or another authority competent to administer oaths, shall be liable to imprisonment of not less than one year.
>
> (2) In less serious cases the penalty shall be imprisonment from six months to five years.

The law is describable as a classical definition, using Boolean operators. The falsity of the testimony is the focus of Section 153, whereas the false oath is the focus of perjury. In both cases, intent is central. Section 15 of the Criminal Code specifies that all criminal laws carry a state of mind requirement of proving intent, unless otherwise specified.

In contrast, the penalties prescribed by the law are not rule-like in this sense. The law instructs courts to determine the sentence in the first instance, and to decide whether the case is "a less serious" one, in which case the court may impose a shorter sentence. This is a classic example of the distinction between rules and standards. Yet what amounts to a standard is merely an expression in a rule that is sufficiently flexible as to give a court significant discretion in deciding how to apply it. As Professor Kim (2007: 413) observes,

> "[T]he inherent uncertainty of legal rules and the need for flexibility to respond to unanticipated situations means that rules cannot definitively determine what a judge should do in every case."

The U.S. perjury statute is quite similar. 18 U.S.C. § 1621(1) reads:

> "Whoever—having taken an oath before a competent tribunal, officer, or person, in any case in which a law of the United States authorizes an oath to be administered, that he will testify, declare, depose, or certify truly, or that any written testimony, declaration, deposition, or certificate by him subscribed, is true, willfully and contrary to such oath states or subscribes any material matter which he does not believe to be true; [...] is guilty of perjury and shall, except as otherwise expressly provided by law, be fined under this title or imprisoned not more than five years, or both. This section is applicable whether the statement or subscription is made within or without the United States."

Unlike the German code, the U.S. version focuses on the defendant's state of mind with respect to the false statement – not with respect to the oath. Yet, no doubt, borderline cases exist, and judges must decide whether such cases come inside or outside the arson law. In particular, judgments about materiality appear in the rule, but require standard-like reasoning.

3.2. Coherence and Pattern Recognition as the Rule of Law[1]

If a person is caught stealing a set of screw drivers from a hardware store, and charged with the theft, it does not feel like any analysis at all is needed. But a great deal of the time, the question arises whether a particular set of facts *should* be considered as coming within a particular legal rule. Often, that requires legal decision makers to decide whether those features to which the event in question is similar to those events already thought to be encompassed within the legal rule are legally significant. The philosopher Nelson Goodman (1972) described the dilemma by noting that we do not judge

[1] The arguments in this section appear in a more expanded form in Solan (2016).

similarity by counting the features that two things have in common, but rather by judging the overall importance of those properties that are shared. He continues:

> "But importance is a highly volatile matter, varying with every shift of context and interest, and quite incapable of supporting the fixed distinctions that philosophers so often seek to rest upon it." (p. 444)

Coherence is a basic rule of law value, adduced by judges and scholars in both the common law and civil law traditions. Below is a statement made by the late Justice Antonin Scalia, an American jurist best known for his adherence to the text and his eschewal of such concerns as the purpose of a statute:

> "Where a statutory term presented to us for the first time is ambiguous, we construe it to contain that permissible meaning which fits most logically and comfortably into the body of both previously and subsequently enacted law. We do so not because that precise accommodative meaning is what the lawmakers must have had in mind (how could an earlier Congress know what a later Congress would enact?), but because it is our role to make sense rather than nonsense out of the *corpus juris*." (West Virginia Univ. Hosps. v. Casey, 499 U.S. 83 [1991]: 100–101, internal citations omitted)

If we do not care what the legislature had in mind, then we must have some other reason for wanting to make sense out of the *corpus juris*. And, of course, we do. Whether or not the legislators had a coherent code in mind, the judges should care, because the most basic rule-of-law values demand that a legal system make sense to the population that it governs. Empirical work bears out this intuition. Recent work by Tom Tyler and his colleagues to the effect that people respond more positively to the legal system when they regard judges as having made decisions based on legitimate concerns lends strong support for this proposition (see Rottman & Tyler, 2014).[2]

The case above from which Scalia is quoted illustrates the tension between legislative primacy and coherence as its own value. In *West Virginia University Hospitals v. Casey* (499 U.S. 83 [1991]), the question was whether a civil rights statute awarding "a reasonable attorney's fee" (42 U.S.C. § 1988) to a prevailing plaintiff included the awarding of the cost of expert fees. Scalia argued coherence on behalf of the majority: "[I]t is our role to make sense rather than nonsense out of the corpus juris" (499 U.S. 83 [1991]: 101).

A number of federal statutes made reference to expert fees, suggesting that statutes intended to include expert fees as part of attorney fees do so expressly (499 U.S. 83 [1991]: 89–90). On the other hand, as the dissenting opinion pointed out, Congress enacted the statute to override a Supreme Court decision that appeared to Congress to be excessively stingy in permitting fee-shifting. It would be surprising if Congress intended to exclude expert fees (499 U.S. 83 [1991]: 113 – Stevens, J., dissenting). The dissent had the last word when the statute was soon amended to include expert fees, once again overriding the Supreme Court (42 U.S.C. § 2000e-5[k] [2012]).

Coherence is deeply embedded in rule-of-law values, as many have noted. For example, Scott Shapiro (2011) bases his theory of "Legality" on the relationship between

[2] My thanks to William Eskridge for pointing out the importance of Tyler's work in this regard.

law-making and interpreting on the one hand, and "plan-making" and execution of plans on the other. In developing this approach, Shapiro notes from the beginning that plans (and thus law-making) must be rational. He comments: "Rationality not only demands that we fill in our plans over time; it also counsels us to settle on plans of actions that are internally consistent and consistent with each other." (2011: 123). This rationality constraint applies generally, both to "bottom-up" planning, the stuff of common law reasoning, and "top-down" planning, the stuff of legislation. Returning to Scalia's two justifications for judges to concern themselves with coherent interpretation, when one interpretation of a law would make it incoherent with the larger body of law and the other would make it fit more rationally, it should be no surprise that judges choose the latter. This will be the case whether because they assume that the enacting legislators would have wanted them to do so, or because, as institutional players, they have an independent obligation to prefer sense to nonsense in statutory interpretation, or both.

Similarly, Dworkin's (1986: 225–275) notion of integrity in law surely incorporates coherence. Dworkin uses the metaphor of each new interpretation of a statute being the equivalent of a new chapter in a chain novel. The interpretation must simultaneously advance the interpretation of the statute to cover (or not cover) new situations consistent with the highest values of the law, and yet be mindful of the statute's past, which includes everything from the societal situation that gave rise to its enactment, including the law's legislative history, to the language of the statute itself, and thus to subsequent interpretations by courts and other institutional actors. Moreover, Dworkin's concept of law as integrity has coherence embedded in its core. For law to have integrity, it must be sufficiently coherent to treat like situations alike.

Scholars writing in the civil law tradition also adduce coherence as an important value in decision making. To a large extent, they also use precedent to demonstrate coherence, although they do so differently, since civil law systems do not have *stare decisis* as a principle of binding law and cases are most often cited for their actual holding. Regardless of the practice concerning citation of precedent, coherence is respected as a legal value in its own right, often under the rubric of "systematic interpretation". Quoting Savigny, Reinhold Zippelius (2008) notes:

> " '[O]nly when we are clear about what a statute's relationship with the overall legal system is, and how the statute is to work within the system,' can we understand the thoughts of the legislator." (61)

He further advocates for coherence as a value in its own right. And Aleksander Peczenik (2008: 230), a legal theorist writing largely in the civil law tradition, bases his entire theory of legal justification on the concept of coherence, linking coherence to rationality, as does Shapiro (2011), trained in the common law tradition.

3.3. Ordinary Meaning: Where the Law Relies on Patterns to Make the Rules Work

Laws are written in words, and the boundaries of word meanings are often not crisp and well-defined. Legal systems tend to operate in a manner that resembles our judgments about category membership and goodness of fit by privileging the "ordinary meaning" of statutory terms (Eskridge, 2016; Slocum, 2015; Solan, 2010). This is no accident, for much of legal reasoning involves making decisions about membership in legally-relevant categories.

To take a classic example, an 1892 U.S. Supreme Court case, *Church of the Holy Trinity v. United States* (143 U.S. 457), construed a law making it a crime to pay for the transportation into the United States of a person performing "labor or service of any kind." The goal of the law was to protect the local labor market (and was probably racist as well). Yet a case was brought against a wealthy church in Manhattan for paying the transportation from London to New York of their new minister. The Court focused on the term "labor," largely ignoring "service." In a famous opinion, a unanimous Court held that the law was not intended to apply to "brain toilers:"

> "It is a familiar rule, that a thing may be within the letter of the statute and yet not within the statute, because not within its spirit, nor within the intention of its makers." (143 U.S. 457 [1892]: 459)

The question, as the Court understood it, was not whether members of the clergy perform "labor or service of any kind." Of course they do. Rather, the question was whether the word "labor" should be applied to the kind of work they do in the context of the statute. To that, the Court answered in the negative. When one uses the word "labor", one thinks of manual labor, not the work of the clergy. In modern legal parlance, the situation in the case was remote from the ordinary sense of the language that the legislature used. Courts no longer talk of the "spirit" of the law as the 1892 Court did, but the analysis has not changed very much.

Whether based on central tendency (prototype analysis) or on fidelity to goal or purpose, this approach to the interpretation of laws is simultaneously rule-like and pattern-like. It is rule-like in that a person did not violate the law unless all of the law's elements were violated. It is pattern-like in that the courts construe the law as applying more readily in core cases than in peripheral ones in meeting the criteria that the statute sets forth. Just because one *can* say that a person has performed labor does not mean that the statute should be applied to that person if the type of work performed seems remote from the goals of the statute.

Many cases in the canon of U.S. cases interpreting statutory law fit this character. Often the court relies on ordinary meaning. In 1919, would a legislature have considered airplanes to be vehicles for purpose of a law banning the removal of stolen vehicles? Such a case came to the U.S. Supreme Court in 1931, *McBoyle v. United States*. Writing for a unanimous court, Justice Oliver Wendell Holmes wrote:

> "No doubt etymologically it is possible to use the word to signify a conveyance working on land, water or air, and sometimes legislation extends the use in that direction, e. g., [illustration omitted]. But in everyday speech 'vehicle' calls up the picture of a thing moving on land." (283 U.S. 25 [1931]: 26).

One could not be clearer in privileging prototypical usage. The Court held that the statute does not apply to stolen airplanes because the mental "picture" that the word evokes does not include them. Holmes did not end the analysis there. In today's world, judges need not explore intuitions about their mental imagery. Instead, they can refer to a corpus of language and determine for themselves whether, as in *McBoyle*, "airplane" collocates with "vehicle," and if so, how frequently it does so compared to other things we call vehicles.

Not only is sticking to ordinary meaning likely to be a good path to fidelity to the legislative will, as Holmes suggests, but it also enhances rule of law values:

> "Although it is not likely that a criminal will carefully consider the text of the law before he murders or steals, it is reasonable that a fair warning should be given to the world in language that the common world will understand, of what the law intends to do if a certain line is passed. To make the warning fair, so far as possible the line should be clear. When a rule of conduct is laid down in words that evoke in the common mind only the picture of vehicles moving on land, the statute should not be extended to aircraft, simply because it may seem to us that a similar policy applies, or upon the speculation that, if the legislature had thought of it, very likely broader words would have been used." (283 U.S. 25 [1931]: 27)

This is an application of the rule of lenity, prevalent in many legal systems, which says that indeterminacy in a criminal statute is to be resolved in favor of the accused. For Holmes, it is at least as important that the law was enacted according to a process that puts people on notice of their obligations as it is that they read the law and know those obligations. A legal system has the right to punish citizens only if it complies with its own legislative obligations.

One need not look back 85 years to find cases that apply the ordinary meaning approach. A law makes it illegal to discriminate against whistle-blowers who disclose corrupt practices within publicly held companies:

> "No [public] company […], or any officer, employee, contractor, subcontractor, or agent of such company, may discharge, demote, suspend, threaten, harass, or in any other manner discriminate against *an employee* in the terms and conditions of employment because of [whistleblowing or other protected activity]." (18 U.S.C. § 1514A [a] [2006])

The issue in *Lawson v. FMR LLC* (134 S. Ct. 1158), decided by the U.S. Supreme Court in 2014 was whether the highlighted term "an employee" in the statute must refer to an employee of the company, or whether it may refer to an employee of a contractor that works for the public company, when it is the contractor's employee who blows the whistle on fraudulent practices in the public company. An investment management firm had fired two people who revealed corrupt practices within a mutual fund whose investments the firm was managing as an outside contractor. The firm argued that the law protects only employees of the public company (the fund in this case) who blow the

whistle on the public company. The Supreme Court disagreed, applying the "ordinary meaning" rule. The most natural way to understand "an employee" in the context of a contractor, the court held, is to construe the term as referring to the contractor's own employee (p. 1165).

At times, the courts agree that they should rely upon a word's ordinary meaning but cannot agree on which of the competing meanings proposed by the parties is the ordinary one. Does a law that makes it a crime to "carry a firearm" "during and in relation to a drug trafficking crime" apply to a person who had illegal drugs in the trunk of his vehicle and a gun in the glove compartment? Or does the law refer only to those who carry guns on their person? (*Muscarello v. U.S.*, 524 U.S. 125 [1998]). A divided court there held that carrying a gun in a car is ordinary enough, and affirmed the conviction. All nine justices agreed that the ordinary meaning should prevail, but they disagreed by a division of five-to-four about which meaning was the ordinary one. The majority decided that what the defendant had done was within the ordinary meaning of the law and affirmed his conviction.

Much of the discussion amounted to an undignified battle among the justices over which dictionary should be considered the most authoritative, and which literary allusions the most representative of ordinary usage. However, in his majority opinion, Justice Breyer also presented a small corpus analysis (pp. 129–130). He indicated that a search using Lexis and Westlaw news libraries revealed that about one-third of the instances of "carry" within a few words of "weapon" involve carrying it in a vehicle.

Mouritsen (2010) demonstrates how, in this case, the use of a linguistic corpus could help to elucidate ordinary usage. Using the Corpus of Contemporary American English ("COCA"), a corpus of more than 500 million words of English from a variety of genres developed at Brigham Young University, Mouritsen showed that the word "carry" is used about six times more frequently to mean "carry on one's person" than to mean "carry in a vehicle." Thus, the majority did not capture the most ordinary sense of the word as it is used in a large corpus of general English. Yet the dispute raises some profound issues. By "ordinary meaning" should the law be concerned with the circumstances in which a word is most commonly used, or should it be concerned with the circumstances in which people are generally comfortable using that word. If the former, Mouritsen's point prevails. If the latter, Justice Breyer's analysis has merit.

Justice Breyer's corpus analysis is not the only instance of judges resorting to big data to determine the ordinary sense of a word or phrase (see Solan & Gales, 2016). Judge Richard Posner, a very prominent appellate judge and legal scholar searched "Google News" in *Costello v. United States* (666 F.3d 1040 [7th Cir. 2012]) to determine whether a woman who invited her boyfriend, who was in the United States without legal immigration papers, was "harboring" him, in violation of a federal law. He found that the verb "harbor" mostly is used in contexts that suggest hiding someone, such as harboring fugitives or harboring Jews. The happy couple in the case at hand, in contrast, were not living in some secret manner, and Posner therefore decided that the

woman had not violated the statute. Two state high courts have also used corpus analysis, both employing COCA. In *State v. Rasabout* (356 P.3d 1258: 1272–73), decided in 2015, the question was whether a gang member who fired twelve bullets from his car as he drove by a house occupied by an enemy of his had "discharged" his gun twelve times, thus committing twelve separate crimes, or whether he discharged it once, by emptying it. The majority opinion relied heavily on dictionaries to reach the conclusion that had discharged the gun twelve separate times and could thus be sentenced accordingly. In a concurring opinion, Associate Chief Justice Thomas Lee turned to COCA, and reached the same conclusion. "Discharge" in the sense of firing a gun is most often used to describe and individual firing of the gun.

Finally, in 2016, in *People v. Harris*,[3] the Supreme Court of Michigan used COCA to examine how the word "information" is ordinarily used. Three police officers stopped a vehicle. One of the officers then assaulted the driver. A passer-by caught the incident on video. At a disciplinary hearing, all three officers lied about what had happened, as later revealed by comparing their testimony to the video. Under Michigan law, police officers are required to testify at disciplinary hearings, but the "information" they give cannot be used against them if any subsequent criminal charges are brought. The law protects their right not to be compelled to incriminate themselves. Because of their false testimony, the officers were charged with obstruction of justice in the disciplinary proceeding. The question was whether their false testimony should be considered "information" or whether misinformation of this sort is outside the scope of the law that would immunize them

The majority on the Michigan Supreme Court held that law does apply, pointing out that COCA contains many examples of people speaking of false or inaccurate information. The dissent had no problem with using COCA, but disagreed with the way the majority conducted its analysis. According to the dissent, when the bare word "information" is used, it virtually always conveys accurate information. Only when it is appropriately modified to signal its falsehood would a hearer or reader conclude that the information is not accurate. In this case, it is not a simple matter to decide what the legislature had in mind: accurate information only, or all uses of the word "information," with whatever modification occurs. Perhaps the principle of lenity should have been applied, resolving the ambiguity in favor of the accused.

Harris provides an important caution in using corpus analysis to determine ordinary meaning: A corpus is nothing more than data. Unless one asks legally-relevant questions, the corpus cannot assist in legal analysis. In *Harris*, the justices on the Supreme Court of Michigan disagreed about what question should be asked of the corpus data, and came to opposite conclusions.

[3] Nos. 149872, 149873, 150042, 2016 Mich. LEXIS 1125 (June 22, 2016).

The disagreement in *Harris* was a linguistic one, but not the only one that sets limits in corpus analysis in statutory interpretation. Courts do not always rule that words must be understood by virtue of the central tendency of their usage. Sometimes they resolve uncertainty in favor of the reading that best furthers a law's goals even if that reading does not conform to the most "ordinary" understanding of the statute's terms. This approach is most consistent with the teleological approach to interpretation, embraced by most Roman law legal systems.

This fact highlights an important consideration in using corpora in legal analysis: The distributional facts that corpora reveal are only useful to the extent that they illuminate distributional facts that the legal system deems relevant. If the judges, consistent with Lynch, Coley & Medin (2000) and Barsalou (1983), determine in a particular case that the purpose of the law is a more important consideration than the prototypical use of the words in the statute, then corpus analysis will not be very helpful. Indeed, judges often concern themselves more with a law's purpose than with which of the competing interpretations is more ordinary.

Despite the focus on ordinary meaning, it is not difficult to find cases in the United States, that focus more on a statute's purpose than on distributional facts about the usage of the words it contains, especially in recent years. Does a law that makes it a crime to destroy financial records, documents, and "tangible objects" to impede a government investigation, enacted to combat financial scandals, apply to a fishing boat captain who threw undersized fish overboard as inspectors began to board his vessel to inspect the cargo? In *Yates v. United States* (574 U.S. ___), decided in 2015, the Supreme Court said no, even though a dead fish is surely a tangible object. The Court decided to pay more attention to the purpose for the statute's enactment than to the ordinary meaning of its terms, much in keeping with the teleological approach.

In another case, *Bond v. United States* (134 S. Ct. 2077 [2014]), a microbiologist who had learned that her husband was the father of the child to whom her best friend was about to give birth took from her place of work a chemical that causes skin irritation and distributed it on her friend's mailbox, door handle, and other such places that her friend was likely to touch. Eventually her friend did come in contact with the chemical, and suffered a mild irritation on her hand, which she treated with warm water. The microbiologist, Bond, was caught doing this mischief and was prosecuted for violating the Chemical Weapons Convention Implementation Act, the statute enacted to implement the Convention on the Prohibition of Chemical Weapons, a treaty to which the United States is a party. That law makes it a crime to make or use a "toxic chemical" except for an approved benign purpose. "Toxic chemical" is defined broadly as "any chemical which through its chemical action on life processes can cause death, temporary incapacitation or permanent harm to humans or animals." The chemical that Bond used, if ingested in large quantities, could cause death or permanent harm to humans, and thus comes within the definition of "chemical weapon" in the statute.

The majority in a divide court would have none of this. To them, this was a local crime that should be handled by the states – not by the law implementing the treaty on chemical weapons. Chief Justice Roberts wrote:

> "The Convention, a product of years of worldwide study, analysis, and multinational negotiation, arose in response to war crimes and acts of terrorism. There is no reason to think the sovereign nations that ratified the Convention were interested in anything like Bond's common law assault.
>
> Even if the treaty does reach that far, nothing prevents Congress from implementing the Convention in the same manner it legislates with respect to innumerable other matters--observing the Constitution's division of responsibility between sovereigns and leaving the prosecution of purely local crimes to the States." (134 S. Ct. 2077 [2014]: 2087, internal references omitted).

One interesting case involves the legal system's treatment of the word "pattern" itself. The U.S. anti-racketeering statute makes it a crime to engage in "a pattern of racketeering activity. Crimes that are considered "racketeering activities" are listed in the statute itself. The law does not fully define the term "pattern," but specifies its meaning to this extent:

> "A 'pattern of racketeering activity' *requires* at least two acts of racketeering activity … the last of which occurred within ten years … after the commission of a prior act of racketeering activity." (18 U.S.C. § 1961[5], emphasis added)

In *H.J., Inc. v. Northwestern Bell Telephone Company* (492 U.S. 229 [1989]), customers sued a telephone company that had given a series of bribes to Minnesota public officials in an effort to obtain a favorable ruling on an application for rate increases that the company sought. Did this effort amount to a "pattern of racketeering activity," or merely an effort to implement a single scheme? The Supreme Court held that the company's activities indeed constituted a pattern: There were numerous bribes within a short period of time, and they were all addressed at accomplishing a particular corrupt result. The Court concluded, "It is this factor of continuity plus relationship which combines to produce a pattern." (p. 239). There was no need to worry about whether the pattern was in service of a single scheme or multiple schemes, according to the Court.

In the U.S., high court decisions interpreting statutes have precedential effect: they must be obeyed by lower courts and, because of the principle of stare decisis, they are not likely to be overturned by the high court itself in a subsequent case. In fact, statutory cases are especially unlikely to be overturned by a subsequent court because the legislature can always decide to override a court decision by simply changing the language of the statute under interpretation. This situation enables us to ask whether there has been a pattern to what the courts call a "pattern" since the Supreme Court decided the issue.

A Lexis database search reveals that since *H.J., Inc.* was decided by the Supreme Court, courts have engaged in more than 300 analyses of "continuity" and "relatedness" to determine RICO liability. Many of these decisions mention the word "pattern" only incidentally. The Supreme Court's effort to create a rule-based approach to deciding

RICO cases by breaking the term "pattern" down into what it believed to be its component parts has not been successful in generating a predictable set of decisions that conform to one's everyday understanding of what constitutes a pattern and what does not.

For example, in the case *Effron v. Embassy Suites* (223 F. 3d 12 [1st Cir. 2000]), an investor in a hotel project in Puerto Rico claimed that the people running construction caused the project to lose money in order to trigger an obligation in the various partnership agreements for the investor to put an additional $1 million into the project. This was done, it was alleged, through a series of seventeen letters and faxes over a 21 month period. The court held that some of these transmissions were not adequately proven and that of the eight that were, almost all had been transmitted within a period of a few months. Thus, the continuity requirement was not met. In another case, *Fleet Credit Corp. v. Sion* (893 F. 2d 441 [1st Cir. 1990]), a company took a secured loan from a bank and the owners then wrote 95 checks over 4.5 years from the company to themselves, leaving the bank with no security. Is this a pattern? The court answered affirmatively, concluding that the continuity requirement had been met, and barely discussing the word "pattern" at all as the operative concept. Together, these two cases illustrate the futility of attempting to create precise lines with respect to concepts that are unspecific by their very nature.

The examples thus far illustrate the fact that although laws are written as hard-and-fast rules, the laws consist of words and word meaning distributes over a conceptual space, forming a pattern. For this reason the need to construe laws, even laws that appear clear on their face, is inescapable. Reliance on context and pragmatic inference (whether about the law's purpose or otherwise) is ubiquitous in legal analysis. Much of the time, of course, the task is not a difficult one. No one would dispute that a stolen car is a stolen vehicle, or that carrying a gun in one's pocket is "carrying a firearm" or that destroying a hard drive that contains incriminating information is destroying a "tangible object" in the context of a financial fraud statute. The hard cases involve situations that appear to be within the outer boundary of the law's language, but remote from prototypical usage, or irrelevant to the law's purpose, or both.

3.4. Laws that Explicitly Call for Pattern-Like Interpretation

Sometimes, in contrast, laws are written to be pattern-like manner as a matter of their drafting. A canon of construction (*ejusdem generis*) tells us that non-exclusive lists are to be limited to the types of things or events that are in the examples that actually occur in the list (see Scalia & Garner, 2012 for detailed discussion). The statute at issue in *McBoyle* (the airplane case discussed above) illustrates this rule. It defines "vehicle" as follows:

> "The term 'motor vehicle' shall include an automobile, automobile truck, automobile wagon, motor cycle, or any other self-propelled vehicle not designed for running on rails" (18 U.S.C. § 408 [1919])

While the definition does not exclude airplanes, the mental model that we form from reading it comfortably includes only vehicles that run along the ground. This principle is grounded in a reasonable folk psychology of language use. The assumption is that when people – including legislators – present examples of what they have in mind and also leave open the possibility of additional items, it is only appropriate to add additional items that are similar to the listed items. Thus, this style of legislation invites statutory interpreters to pattern the statute around a loosely structured set of criteria.

Laws are also sometimes written to describe the prototypical case, allowing for deviation when the ordinary situation does not obtain. The United States Code contains more than 400 such provisions and state codes contain many more in the aggregate.[4] Typical examples include:

> "A term of probation commences on the day that the sentence of probation is imposed, *unless otherwise* ordered by the court." (18 U.S.C. § 3564[a])

> "*Unless otherwise* stated, the requirements applicable to cigarettes under this chapter [21 USCS §§ 387 et seq.] shall also apply to cigarette tobacco." (21 U.S.C. § 387)

> "Definitions. For purposes of this section, *unless otherwise* provided or indicated by the context—
> (1) the term 'Administration for Community Living' means the Administration for Community Living of the Department of Health and Human Services;
> (2) the term 'Federal agency' has the meaning given to the term 'agency' by section 551(1) of title 5, United States Code" (42 U.S.C. § 3515e)

In some instances, the legislature places limits on the legitimacy of straying from the prototype:

> "Hearings under this section shall not be public, *unless otherwise* ordered by the Board *for good cause shown*, with the consent of the parties to such hearing." (15 U.S.C. § 7215)

It is worth noting that these data are also distributional: they reflect the distribution of particular linguistic practices within the legislature. Moreover, the search for such examples constitutes a corpus analysis in its own right. Here, however, the goal of the analysis is not to determine how commonplace the language of a legal document is by comparing the usage in the document to the way the language is used in general speech and writing. Rather, the corpus used here is a legal corpus, with the goal of determining whether that corpus contains a particular linguistic structure, and if so, how often.

4. Patterns in Legal Language

It has been observed that legal language has its own characteristics, many of which are facts about distribution. For example, legal language has its own vocabulary, some of

[4] A Lexis search conducted 28 October 2016 of U.S. Code library "unless otherwise" yielded 423 hits. The same search among state codes yielded 577 hits.

which sounds arcane, and some of which is nothing more than the specialized glossary of a field (see Goźdź-Roszkowski, 2011; Tiersma, 1999; Mattila, 2006). Especially interesting are words that are in common usage outside the legal sphere, but which have specialized legal meaning as well (see Schauer, 2015). To take a classic example, "consideration" is a technical term in contract law meaning the thing given in exchange as part of a bargain, and also has an everyday meaning, thoughtful contemplation.

Similarly, framing effects are important in legal argumentation. The lawyer representing an individual in a deportation hearing is more likely to refer to her client as an undocumented worker, the while the government lawyer may speak of an illegal alien. Moreover, different legal genres may show preferences for different vocabulary: contracts and statutes, while both authoritative legal documents, do not look or sound the same. By the same token, in a rape case in which a man is accused of raping a woman, the prosecutor is more likely to refer to the woman as "the victim," the defense lawyer to refer to her as "the complaining witness." In one publicized case, a judge ordered the prosecutor to stop referring to the woman as "the victim." The defendant was later acquitted. Legal anthropologists have described in considerable detail the ways in which word choice frames issues and the ways in which lawyers attempt to make choices to direct a case's vocabulary in a direction beneficial to the lawyer's client. (see, e.g., Matoesian, 1993; Conley & O'Barr, 1998).

To illustrate, in the United States, "undocumented immigrants" and "illegal aliens" refer to the same groups of people but have very different connotations. Nuñez (2013) conducted a corpus analysis of "alien," "immigrant," and "citizen," using Brigham Young University's Corpus of Contemporary American English (COCA). In order to focus her study on uses of "alien" that do not involve extra-terrestrials and the like, she searched for words that appear in close proximity to both "alien" and "immigrant" (collocates that the two have in common) and compared the relative frequency of occurrence, a design that is likely to capture the intended sense of "alien," at least most of the time. She found that "criminal" and "illegal" occur more frequently with "alien" than with "immigrant," and that "new," "legal," "undocumented," "American" and "other" appear more frequently with "immigrant" than with "alien." Thus, we speak of "undocumented immigrants," and "illegal aliens."

Similarly, Gales (2009) has demonstrated that congressional debate about "diversity" in the context of immigration law reform is replete with negative markers, suggesting very mixed feelings about adjusting immigration policy to promote diversity, even among those who purport to support such initiatives. In other instances, such analysis can differentiate among the genres with which legal actors express themselves. For example, contracts and legislation, while both authoritative legal documents, do not typically use the same vocabulary or syntactic style (Goźdź-Roszkowski, 2011).

These are important issues that have received a great deal of attention. I devote little spaced to them here not because I believe them to be unimportant, but rather because the focus of this article is on interpretive issues.

5. Patterns in the Law's Application

By the same token, the application of laws may reveal patterns – not all of them reflecting positively on a particular legal system. Because these patterns do not necessarily involve issues of language, I will comment on them only briefly here.

To take one example, studies in the United States show that the death penalty is meted out disproportionally to offenders convicted of killing white victims. In 1987, the United States Supreme Court held in *McCleskey v. Kemp* (481 U.S. 279 [1987]) that studies demonstrating that the death penalty is not applied evenly in the state of Georgia were not sufficient to lead to the reversal of McCleskey's death sentence for armed robbery and murder because the studies could not demonstrate that McCleskey himself received an unfair trial. Thus, the murder and capital punishment laws are written as rules, but the application of the law formed a pattern that reflected racism in the legal system.

Such problems are not linguistic in nature except to the extent that they involve the construal of laws to license the conduct. Surely, however, such practices as racial profiling in policing the highways, and the more gruesome example involving the death penalty are not about battles over the interpretation of particular laws, even if in both instances the law is applied properly to those who are prosecuted, but otherwise discriminatorily.

6. Conclusion

The principal goal of this article has been to illustrate how much statutory analysis in law is based on the notions of central tendency and goal orientation, two considerations that do not fit into rule-based analysis. When the law chooses to privilege central tendency, generally called "ordinary meaning" in legal contexts, and prototype in linguistics, corpus analysis can be useful, as both scholars and judges have recognized. Corpus analysis cannot solve all of the legal system's interpretive puzzles. But when the legal system commits to rendering judgments based on the kind of information that a corpus contains, use of the corpus is far superior than hoping for judges to resolve conflicting information about the distribution of meaning across a population.

DOI: 10.14762/jll.2017.046

References

Armstrong, Sharon L., Gleitman, Lilsa R. & Gleitman, Henry (1983). What Some Concepts Might Not Be. *Cognition, 13*, 263–308. DOI: 10.1016/0010-0277(83)90012-4.

Barsalou, Lawrence W. (1983). Ad hoc categories. *Memory & Cognition, 11*(3), 211–227. DOI: 10.3758/BF03196968.

Chomsky, Noam (2005). Three factors in language design. *Linguistic Inquiry, 36*, 1–22. DOI: 10.1162/0024389052993655.

Conley, John M. & O'Barr, William M. (1998). *Just Words: Law, Language, and Power*. Chicago: University of Chicago Press.

Conrad, Susan M. & Biber, Douglas (2004). The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica: International Annual for Lexicography, 20*, 56–71. DOI: 10.1515/9783484604674.56.

Dworkin, Ronald (1986). *Law's Empire*. Cambridge, MA: Harvard University Press.

Eskridge, William N., Jr. (2016). *Interpreting Law: A Primer on How to Read Statutes and the Constitution*. St. Paul, MN: Foundation Press.

Gales, Tammy (2009). Diversity as enacted in U.S. policy and law: A corpus-based approach. *Discourse & Society, 20*(2), 223–240. DOI: 10.1177/0957926508099003.

Goldberg, Adele & Jackendoff, Ray (2004). The English resultative as a family of constructions. *Language, 80*, 532–568.

Goodman, Nelson (1972). Seven Strictures on Similarity. In Goodman (Ed.), *Projects and Problems* (pp. 437–446). New York: Bobbs-Merrill Company.

Goźdź-Roszkowski, Stanisław (2011). *Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study*. Frankfurt: Peter Lang.

Jackendoff, Ray (2003). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Cambridge, MA: MIT Press.

Jackendoff, Ray (2008). Construction After Construction and its Theoretical Challenges. *Language, 84*, 8–28. DOI: 10.1353/lan.2008.0058.

Kim, Pauline T. (2007). Lower court discretion. *NYU Law Review, 82*, 383–442. Available at nyulawreview.org/sites/default/files/pdf/NYULawReview-82-2-Kim.pdf.

Lynch, Elizabeth B., Coley, John D. & Medin, Douglas L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition, 28*, 41–50. DOI: 10.3758/BF03211575.

Matoesian, Gregory M. (1993). *Reproducing Rape: Domination through Talk in the Courtroom*. Chicago: University of Chicago Press.

Mattila, Heikki E.S. (2006). *Comparative Legal Linguistics*. Aldershot: Ashgate.

Mouritsen, Stephen C. (2010). The Dictionary Is Not a Fortress: Definitional fallacies and a corpus-based approach to plain meaning. *Brigham Young University Law Review, 2010*, 1915–1978. Available at digitalcommons.law.byu.edu/lawreview/vol2010/iss5/10.

Mouritsen, Stephen C. (2011). Hard Cases and Hard Data: Assessing corpus linguistics as an empirical path to plain meaning. *Columbia Science and Technology Law Review, 13*, 156–205. Available at stlr.org/cite.cgi?volume=13&article=4.

Nuñez, D. Carolina (2013). War of the words: Aliens, immigrants, citizens, and the languge of exclusion. *Brigham Young University Law Review, 2013*, 1517–1562. Available at digitalcommons.law.byu.edu/lawreview/vol2013/iss6/9.

Peczenik, Aleksander (2008). *On Law and Reason*. Dordrecht: Springer.

Pinker, Steven (1999). *Words and Rules: The Ingredients of Language*. New York: Basic Books.

Rosch, Eleanor (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233.

Rottman, Dan B. & Tyler, Tom R. (2014). Thinking about Judges and Judicial Performance: Perspective of the public and court users. *Oñati Socio-Legal Series*, *5*, 1046–1070. Available at opo.iisj.net/index.php/osls/article/view/343.

Scalia, Antonin & Garner, Bryan (2012). *Reading Law: The Interpretation of Legal Texts*. St. Paul, MN: Thompson/West.

Schauer, Frederick (Ed.) (2009). *Thinking Like a Lawyer*. Cambridge, MA: Harvard University Press.

Schauer, Frederick (2015). On the relationship between legal and ordinary language. In Solan, Ainsworth & Shuy (Eds.). *Speaking of Language and Law: Conversations on the Work of Peter Tiersma* (pp. 35–38). New York: Oxford University Press.

Shapiro, Scott (2011). *Legality*. Cambridge, MA: Harvard University Press.

Shuy, Roger (Ed.) (2015). *The Language of Fraud Cases*. New York: Oxford University Press.

Slocum, Brian G. (2015). *Ordinary Meaning: A Theory of the Most Fundamental Principle of Legal Interpretation*. Chicago: University of Chicago Press.

Solan, Lawrence M. (2009). Linguistic knowledge and legal interpretation: What goes right, What goes wrong. In Schauer (Ed.), *Thinking Like a Lawyer* (pp. 66–87). Cambridge, MA: Harvard University Press.

Solan, Lawrence M. (2010). *The Language of Statutes: Laws and their Interpretation*. Chicago: University of Chicago Press.

Solan, Lawrence M. (2016). Precedent in Statutory Interpretation. *North Carolina Law Review* 94, 1165–1234. Available at scholarship.law.unc.edu/nclr/vol94/iss4/2.

Solan, Lawrence M. (2017), Linguistic knowledge and legal interpretation – what goes right, what goes wrong. In Slocum (Ed.). *The Nature of Legal Interpretation: What Jurists Can Learn about Legal Interpretation from Linguistics and Philosophy* (pp. 66–87). Chicago: University of Chicago Press.

Solan, Lawrence M. & Gales, Tammy (2016). Finding ordinary meaning in law: The judge, the dictionary or the corpus. *International Journal of Legal Discourse*, *1(2)*, 253–276. DOI: 10.1515/ijld-2016-0016.

Solan, Lawrence M., Rosenblatt, Terri & Osherson, Daniel (2008). False consensus bias in contract interpretation. *Columbia Law Review*, *108*, 1268–1300. Available at jstor.org/stable/40041799.

Tiersma, Peter (1999). *Legal Language*. Chicago: University of Chicago Press.

Vogel, Friedemann, Hamann, Hanjo & Gauer, Isabelle (2017). Computer-Assisted Legal Linguistics. Corpus Analysis as a New Tool for Legal Studies, *Law & Social Inquiry*, *42*, early view. DOI: 10.1111/lsi.12305.

Yang, Charles (2002). *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

Zippelius, Reinhold (2008). *Introduction to German Legal Methods*. Durham, NC: Carolina Academic Press.

International Journal of
Language & Law

Mouritsen, JLL 6 (2017): 67–89
www.languageandlaw.de

Corpus Linguistics in Legal Interpretation

— An Evolving Interpretative Framework

Stephen C. Mouritsen[*]

Abstract

When called upon to interpret the undefined words in a legal text, U.S. judges will often invoke a rule (or canon) of interpretation called the "plain meaning rule," which holds that if the language of the text is clear and unambiguous, courts cannot consider any extrinsic evidence to determine what the text means. But U.S. courts have no uniform definition of what "plain meaning" actually means and no systematic method for discovering and resolving ambiguities in legal texts. Faced with these challenges, some U.S. judges and academics have recently begun to consider the use of corpus linguistics to resolve uncertainties in the interpretation of legal texts. A corpus-based approach to legal interpretation promises to increase the objectivity and predictability of decisions about the meanings of legal texts. However, such an approach also presents a number of theoretical problems that must be addressed before corpus methods can be fully incorporated into a theory of legal interpretation. This article documents this recent turn to corpus linguistics in legal interpretation and outlines some of the challenges facing the corpus-based approach to legal interpretation.

Keywords

corpus linguistics, statutory interpretation, legal interpretation, plain meaning, ordinary meaning, legal linguistics

Submitted: 14 February 2017, accepted: 10 August 2017, published online: 4 September 2017

[*] Associate at the University of Chicago Law School and Adjunct Professor of Law and Corpus Linguistics at Brigham Young University, USA, stephen.mouritsen@uchicago.edu. The author wishes to express his gratitude to the JLL editors and the two anonymous reviewers for their helpful comments on an earlier draft of this article.

1. Introduction

Judges and lawyers are often presented with problems of interpretative uncertainty – ambiguous legal texts that present two or more potential interpretations or vague legal language with a range of possible meanings. When faced with such interpretative challenges, jurists often look for guidance in statutory definitions or prior cases addressing similar statutory language.[1] Where the relevant statutory terms are undefined, or where no settled ruling governs the interpretative outcome, jurists are left to cast about for other interpretive heuristics. Often, jurists must attempt to resolve questions of interpretive uncertainty by relying on their linguistic intuition. And, increasingly in the U.S. jurisprudence, judges are appealing to general-use dictionaries to resolve questions of interpretive uncertainty (Brudney & Baum, 2013: 495; Thumma & Kirchmeier, 1999: 248–260; Thumma & Kirchmeier, 2010: 77; Note, 1993–1994: 1454 had even showed a nearly exponential increase in the Court's reliance upon dictionaries). But human linguistic intuition is at best a problematic guide to the predictable and objective resolution of interpretative uncertainty in legal texts.[2]

Human decision making is subject to a host of well-documented cognitive biases that may affect objectivity (Sunstein, 1997: 1176), and a great deal of objective linguistic information is not available through introspection (McEnery & Wilson, 2001). Moreover, dictionaries, whatever their merits, rarely contain the answers to the interpretative questions for which they are cited in U.S. courts. While the general-use dictionaries often cited by U.S. courts attempt to document the range of possible meanings of a given word, they cannot be relied upon to show the meaning of a given word in a given statutory context: "A dictionary, it is vital to observe, never says what meaning a word must bear in a particular context. Nor does it ever purport to say this." (Hart Jr. & Sacks, 1994: 1190).

Recognizing this problem, a few U.S. courts and academics have begun to consider the use of corpus linguistics to resolve uncertainties in the interpretation of legal texts. A corpus-based approach to legal interpretation promises to increase the objectivity and predictability of decisions about the meanings of legal texts. However, such an approach also presents a number of theoretical problems that must be addressed before corpus methods can be fully incorporated into a theory of legal interpretation.

[1] Eskridge Jr. (2016: 74) described the "statutory definition canon" as follows: "When a statute defines a word or phrase, interpreters should follow the ordinary meaning of the statutory definition", and notes (139) that "future applications of statutory law to newer facts will not only consider the plain meaning and whole act, but will also (and should) consider precedents interpreting relevant statutory provision."

[2] For example, inter-annotator agreement on fine-grained Word Sense Disambiguation ("WSD") tasks is often poor (Véronis, 1998). The task of determining which of two competing, fine-grained senses of a given word is appropriate in a given context is often similar to the task faced by a judge in interpreting a vague or ambiguous statutory directive.

Set forth below is a brief discussion of the emergence of the corpus-based approach to legal interpretation in U.S. jurisprudence, as well as a discussion of a number of the challenges facing the corpus-based approach to legal interpretation.

2. Prior Use of Linguistic Corpora in a Legal Context

Until very recently in U.S. courtrooms, the use of linguistic corpora in has been the domain of experts. For example, in the case of *LG Electronics USA, Inc. v. Whirlpool Corp.*, LG Electronics USA, Inc. ("LG"), an electronics manufacturer, sued its competitor Whirlpool Corporation ("Whirlpool") for false advertising (661 F.Supp.2d 940 [2009]). LG manufactured a clothing dryer called a Tromm Steam Dryer. The dryer injected steam into the dryer drum in order to reduce wrinkles (id.: 943–944). The water was heated to a boil in an attached boiler and then injected into the dryer drum. Whirlpool began to market a competing "Steam Dryers" (id.: 943). Rather than produce steam through boiling, the Whirlpool Steam Dryers simply injected water into the dryer drum during the drying processes. The water would vaporize when it came in contact with the heated clothing. The case then turned in large measure on the meaning of the word *steam* (id.: 945–946). Linguist Judith Levi submitted an expert report in which she analyzed the different uses of the noun *steam* data from an electronic database (Levi, 2008, using the Westlaw ALLNEWS and USNEWS databases). Levi found numerous examples of steam in which steam was used to mean visible water vapor that can be observed at room temperature. Whirlpool would ultimately prevail in the suit.

In another case, Microsoft sued Apple to try to prevent Apple from registering the phrase "app store" as a trademark.[3] In that case, linguist Robert A. Leonard analyzed evidence from the Corpus of Contemporary American English ("COCA") and concluded that "the predominant usage of the term APP STORE is as a proper noun to refer to Apple's online application marketplace" (Leonard, 2008).

These uses of linguistic corpora by experts fit into a familiar pattern of the use of linguistic experts in U.S. product and trademark cases.[4] While the use of corpus data in such cases is comparatively new, by keeping the corpus data in the hands of the expert, such cases do not upset the existing paradigm of having data-driven linguistic data enter the courtroom through experts. Increasingly, however, judges and lawyers are departing from this traditional paradigm, performing their own corpus linguistic analysis. Not only do these cases represent a change in the paradigm because judges

[3] In the Matter of Application Serial No. 77/525,433 (July 17, 2008).

[4] Of course, product and trademark cases are not the only cases in which corpus data is used by experts in U.S. courts. Corpus linguistics can play an important role in questions of author identification (Kredens & Coulthard, 2012), and corpus-based techniques form an important part of the document discovery process where electronically stored documents are concerned (Hietala Jr., 2014: 603).

and lawyers are accessing sources of empirical research directly, but because they are aimed at entirely different questions. Experts called in to testify in cases like *LG Electronics* and the *App Store* case are asked to opine about public perception of a mark that was prepared by non-lawyer designers and marketing professionals in order to influence the perceptions of the lay public. As we will see below, the paradigm is entirely different when a text prepared in what is ostensibly specialized, legal language is interpreted by a professional class of lawyers and judges. This raises the question about whether or not a corpus comprised of non-legal texts can be used effectively to interpret a legal text. We discuss this problem below.

3. Quasi-Corpora and the Data Impulse

It is perhaps unsurprising that U.S. judges who routinely rely on sophisticated, heavily annotated databases of case law, rules, and statutes, and who undoubtedly – like most other members of contemporary society – routinely turn to the Internet for answers to quotidian questions, would eventually begin to turn to electronic data when attempting to resolve questions of legal interpretation.

Before the advent of the personal computer (and even today), case law from the numerous state and federal courts in the United States was published in bound volumes called "reporters" and then sorted into topical indices called "digests" (*e.g.*, the West American Digest System – West, 1909: 4). The digest was a printed index in which an attorney would search for a given topic (*e.g.*, breach of contract, the rule against perpetuities), trusting that the human annotator who had prepared the digest had properly indexed all of the relevant case law from the jurisdiction in question. However, because of the sheer volume of precedent produced by the numerous state and federal courts each year, commentators began to express concern that the human annotators charged with indexing the nation's case law would be overwhelmed by the number of cases to index and would not be able to capture all of the relevant precedent for a given topic. It was estimated, for example, that as early as 1961 "there were 2.2 million reported cases (this figure was increasing at a rate of 25,000 per year), […] and 2 million entries in descriptive word indices" (Note, 1967: 993, citing Dickerson, 1961: 902). This immense volume of case law, when paired with the imperfect performance of human annotators, meant that "the element of chance" necessarily played "an increasingly significant role in the locating of relevant information" (Note, 1967: 993). As one early commentator noted:

> "There is strong suspicion that the mountain of precedents has grown to such size that legal research ordinarily consists of no more than snatching the first bit of relevant material that can be found and then flying by the seat of the pants. Let us not delude ourselves. Our legal system depends on precedent to insure that we have a government of laws and not of men, but in practice we rely more on gen-

> eralized experience, on the lawyer's 'feel' based on vague personal recollections of precedent, rather than on precedent itself." (Melton & Bensing, 1961: 248)

This ever-expanding "mountain of precedent" and the concern about human annotators' inability to properly index the same (together with the rise in computing power over the last half of a century) led to the development of the sophisticated commercial legal research databases that U.S. lawyers now rely on every day (*e.g.*, Westlaw, Lexis, Bloomberg Law). While some have expressed concern that the use of computers in legal research dulls lawyers' legal reasoning ability (e.g., Bintliff, 1996: 339; Lien, 1998: 85–86), today nearly every U.S. judge's chambers and nearly every U.S. lawyer's office has a personal computer that links to an online repository of millions of cases, statutes, and legal rules. Lawyers, even those who otherwise lack sophisticated knowledge of computers, are nevertheless able to perform complex Boolean searches to locate every case, statute, or rule, addressing a given topic, in a given jurisdiction. As was predicted more than half a century ago, the computer has not altogether replaced the lawyer in performing legal research: "the lawyer will still have to analyze and the judge will still have to decide" (Note, 1967: 993). However, the use of such computational research databases can both reduce the amount of time a lawyer spends in conducting research[5] and increase the lawyers' certainty in the completeness of those results:

Similarly, judges and lawyers like almost every other member of contemporary society naturally rely on the Internet to answer everyday questions. More controversially, many judges have been unable to resist the impulse to conduct factual research using Internet searches. As Judge Richard A. Posner has recently observed:

> "The Internet […] ha[s] made it much easier for judges to conduct their own factual research […] rather than having to rely entirely on what the lawyers serve up to them. And because it is easier, judges (and their law clerks) are doing more of it, and this has given rise to controversy." (Posner, 2013: 134; see also Thornburg, 2008: 131)

Because judges and lawyers already appeal to curated, commercial legal databases to look for legal rules and precedent, and because judges and lawyers have a natural impulse to look for answers to questions using Internet searches, it is not surprising that judges might turn to either of these sources in order to attempt to resolve questions of legal interpretation.

For example, in the case of *Muscarello v. United States*, the United States Supreme Court was called upon to interpret the phrase *carries a firearm* from the Omnibus Crime Control and Safe Streets Act of 1968 (later codified as 18 U.S.C. § 924[c][1]) and to determine whether Congress intended by that term to include the notion of *conveyance in a vehicle* (524 U.S. 125 [1998]: 129, discussed in Mouritsen, 2010: 1915). *Muscarello* is a ground-breaking case because it is the first case in which a court relied on a quantita-

[5] See Melton & Bensing (1961: 248): "The computer performs repetitive, routine tasks more thoroughly, at lower cost, and faster than human beings. Computers therefore can relieve the human being of such tasks and allow him to devote his full energies and time to the reasoning tasks which he, of course, performs far better than a computer."

tive analysis of linguistic data to address a question of statutory interpretation. Writing for the majority, Justice Breyer stated that

> "to make certain that there is no special ordinary English restriction (unmentioned in dictionaries) upon the use of 'carry' […] we have surveyed modern press usage, albeit crudely, by searching computerized newspaper data bases." (524 U.S. 125 [1998]: 129)

These searches were conducted in a New York Times database found in Lexis/Nexis, and a U.S. News database found in Westlaw. Justice Breyer then describes the search parameters and results as follows:

> "We looked for sentences in which the words 'carry,' 'vehicle,' and 'weapon' (or variations thereof) all appear. We found thousands of such sentences, and random sampling suggests that many, perhaps more than one-third, are sentences used to convey the meaning at issue here, i.e., the carrying of guns in a car." (524 U.S. 125 [1998]: 129)

The key flaw in the *Muscarello* court's attempt at a sort of quasi-corpus linguistic search is found in its search parameters. If the court wants to know whether the phrase *carries a firearm* ordinarily includes the notion of conveyance in a *vehicle*, then the search cannot contain the word *vehicle*. Justice Breyer should have examined sentences that contained references to "carry" and "firearm" and determined how many referred to conveyance in a vehicle versus conveyance on one's person.

A similarly approach was taken in *United States v. Costello*. In that case, the Seventh Circuit Court of Appeals (666 F.3d 1040 [2012]: 1041–1042) was asked to determine the meaning of harboring in the context of an statute which imposes an enhanced prison sentence of five additional years upon anyone who "knowing […] the fact that an alien has come to, entered, or remains in the United States in violation of law, conceals, harbors or shields from detection […] such alien" (8 U.S.C. § 1324[a][1][A][iii]).

The defendant was an American citizen charged with harboring her boyfriend, whom she knew to have entered the United States unlawfully. (666 F.3d 1040 [2012]: 1042 – the boyfriend is not named in the opinion and is instead referred to as "the boyfriend".) The two had lived together for about a year, until the boyfriend was arrested on a federal drug charge, spent several years in prison, and was then sent back to Mexico. The boyfriend returned to the United States and upon arrival, called Ms. Costello and requested a ride from the bus station and resumed residing with Ms. Costello. There was no evidence that Ms. Costello attempted to conceal her boyfriend from the authorities – only that she offered him a place to stay.

The government cited a dictionary to argue that harbor meant merely *to shelter*. But both senses of the verb *harbor* at issue in the case are attested in dictionaries. *Harbor* can mean either "to give shelter or refuge to" (*see* Webster's Third New International Dictionary, sense 1a(1) of *harbor*) or "to receive clandestinely and conceal" (*see* Webster's Third New International Dictionary, sense 1a(2) of *harbor*). Judge Posner acknowledges at least one problem with respect to relying on dictionaries, noting that "[d]ictionary

definitions are acontextual, whereas the meaning of sentences depends critically on context, including all sorts of background understandings." (id.)

Rather than dwell on dictionary definitions, Judge Posner engages in what may be the first attempt by a judge to justify the interpretation of a statute with by means of a search in the Google search engine. Judge Posner states: "A Google search […] of several terms in which the word 'harboring' appears – a search based on the supposition that the number of hits per term is a rough index of the frequency of its use – reveals the following […]" Judge Posner then lists the results of searches for a number of phrases that include the word *harboring*, including *harboring fugitives, enemies, refugees, victims, flood victims, victims of disasters, victims of persecution, guests, friends, Quakers,* and *Jews* (id.). Judge Posner concludes that

> "[i]t is apparent from these results that 'harboring,' as the word is actually used, has a connotation – which 'sheltering,' and a fortiori 'giving a person a place to stay' – does not, of deliberately safeguarding members of a specified group from the authorities, whether through concealment, movement to a safe location, or physical protection." (id.)

There are a number of reasons why Google might appear at first blush to be a good source for data-driven analysis of language usage.

> "The web is enormous, free, immediately available, and largely linguistic. As we discover, on ever more fronts, that language analysis and generation benefit from big data, so it becomes appealing to use the web as a data source." (Kilgarriff, 2007: 147)

As the world's most popular, freely available online search engine, Google has no entry costs and has a familiar, easy-to-use interface. It is hard to imagine a judge's chambers or law office that does not have access to Google.

But the notion that citation to Google could provide even a "rough index of the frequency of [a term's] use" (666 F.3d 1040 [2012]: 1042) is so beset with methodological problems that it renders the results, if not entirely arbitrary, then at least deeply problematic. For example, Judge Posner examines the comparative hit counts of a number of words as they co-occur with *harboring*, but never explains how he came up with the list of words in question. The opinion does not provide any sort of selection criteria for the nouns included in the search, nor does it explain whether or not any additional word pairings were examined but not included. We are left with the impression that Judge Posner's choice of these words was based on his own linguistic intuition. Judge Posner examines eleven words or phrases: *fugitives, enemies, refugees, flood victims, victims of disasters, victims of persecution, guests, friends, Quakers, Jews.* (For reasons not explained, Judge Posner excludes the statutory term itself: *alien*.) Of the eleven words or phrases examined by Judge Posner, only *fugitives* and *Jews* appears.

A Google search offers no lemmatization or grammatical tagging, that is, Google does not offer an easy way to search for the verb *to harbor* but not the noun *harbor* in a single search (O'Keeffe & McCarthy, 2010: 172). The words in a corpus like the COCA, which have been automatically labeled with meta-data related to part-of-speech, so

that a search for the verb *harbor* can easily be tailored reveal only the verbal form of harbor, with all of its potential inflections. In addition, Judge Posner's searches ignore the morphology of the words in his searches. In order to perform a set of searches that even begins to account for the most rudimentary range of the potential uses of *harbor* in the phrases the *Costello* opinion examines, we would have to perform 132 separate Google searches. These searches would include four verb forms (*harbor, harbors, harboring, harbored*) multiplied by three noun forms (e.g., *a fugitive, the fugitive, fugitives*) multiplied by the eleven separate phrases examined in the opinion. And this would not even begin to account for the variety of words that might intervene between the verb *harbor* and its nominal object.

Google cannot meaningfully be said to represent any particular speech community.[6] A single, English language search in Google may represent speech from a wide variety of language users, *e.g.*, the Times of India (timesofindia.indiatimes.com) or the Ghanaian Times (ghanaiantimes.com.gh) – both English language papers from presumably different dialect regions. We have no reliable way of knowing what these searches contain. Google searches

> "are sorted according to a complex and unknown algorithm (with full listings of all results usually not permitted) so we do not know what biases are being introduced. If we wish to investigate the biases, the area we become expert in is googleology not linguistics." (Kilgarriff, 2007: 148)

A more fundamental problem with Judge Posner's use of Google is that the Google hit counts are notoriously unreliable, as they are based on the number of webpages with a given word, not the number of times a given word occurs. Google hit returns can vary by geography, by time of day and day after day. In one experiment,

> "queries repeated the following day gave counts over 10% different 9 times in 30 [...] The reasons are that queries are sent to different computers, at different points in the update cycle, and with different data in their caches." (Kilgarriff, 2007: 148)

While Justice Breyer's news database approach in *Muscarello* and Judge Posner's Google-based approach in *Costello* have numerous flaws, one of the chief benefits of their respective approaches is that their flaws are visible. Rather than merely declare a particular sense of a word to be the ordinary meaning based on their respective intuitions, Justice Breyer and Judge Posner have each performed a flawed experiment, but the experiments are, at the very least, replicable and falsifiable.

In addition, both cases demonstrate two key facts that may lead to an increase in the use of empirical methods for legal interpretation. First, both cases demonstrate a recognition of the inadequacy of existing tools to resolve questions of interpretation. In both *Muscarello* and *Costello*, the parties and the judges cite dictionary definitions to support their interpretation of the relevant statutes and in both cases, citing to dic-

[6] Dickerson (1983: 1154) defines "speech community" as "simply a group of people who share a common language (or sublanguage) and thus a common culture (or subculture), which in turn defines the context that conditions the utterances that occur within it."

tionaries fails to eliminate the ambiguity in the texts or reveal the texts' ordinary meaning. Second, in both cases, the judges (likely recognizing the inadequacy of a dictionary-based approach) gave way to a contemporary impulse to look for answers in easily available data through a news search and a Google search respectively. While we may take exception to both the methods and the sources relied upon in these opinions, these opinions demonstrate that the impulse to replace dictionaries with readily available language data will become harder and harder for judges and lawyers to ignore. The best course may be to ensure that these judges and lawyers have access to the best available sources of language data, and have training in the best linguistic methods for investigating meaning.

4. Corpus Linguistics in Statutory Interpretation

While early attempts at a data-driven approach to statutory interpretation were innovative, they suffered from a number of methodological problems – problems that could be addressed with the use of sophisticated annotated corpora. In the fall of 2010, two documents, a law review article and an amicus brief were published setting forth similar corpus-based approaches to statutory interpretation.[7]

4.1. FCC v. AT&T

In the case of *FCC v. AT&T* (131 S. Ct. 1177 [2011]), the United States Supreme Court was asked to determine whether the "personal privacy" exemption of the Freedom of Information Act ("FOIA"), 5 U.S.C. § 552(b)(7)(C), applies to corporations. Rather than rely on "scattershot, impressionistic evidence" like dictionary definitions, or their own linguistic intuitions, the justices instead "drew on some nuanced linguistic expertise" to determine the scope of FOIA's "personal privacy" exemption (Zimmer, 2011).[8] The brief, written by attorney Neal Goldfarb and submitted on behalf of the Project for Government Oversight, used collocation data to show that the documented usage of the adjective "personal" could not sustain an interpretation of FIOA's "personal privacy" exemption that would apply that term to corporations.[9] The brief examines data from three large linguistic corpora to demonstrate that "*personal* has developed a specialized meaning such that it is used with regard to human beings, not corporations"

[7] For a more detailed discussion of the interpretative problems in *Costello* and *Muscarello*, and a corpus-based approach to resolving these interpretive problems, see Lee & Mouritsen (forthcoming 2017).

[8] Ben Zimmer is the former *On Language* columnist for the New York Times and language columnist for the Atlantic; he now writes for Wall Street Journal.

[9] Brief for the Project on Government Oversight et al. as Amici Curiae Supporting Petitioners, FCC v. AT&T Inc., No. 09-1279 (U.S. Nov. 16, 2010).

(16). The analysis proceeds by "querying each corpus so that it returns the nouns that appear most frequently in the position immediately following *personal*" (16). In virtually every case, the brief concludes, the nouns found paired with the adjective "personal" were those that made exclusive reference to human beings. These included *personal life*, *personal experience*, *personal relationship*, *personal friend*, and *personal question* (17).

The results of Goldfarb's query have a number of immediate advantages over the searches performed by Judge Posner in the *Costello* opinion. To begin with, Goldfarb searched a principled corpus of American usage, designed to sample the native speech of the speech community intended to be governed by FOIA's provisions. Goldfarb has relied on the corpus interface, and not his own intuition, in order to generate his list of collocations. And while Goldfarb does not list the statistical frequency of these collocations, it would have been easy for him to do so – ranking them from most statistically frequent to least. Indeed we can easily duplicate both Goldfarb's results and his methodology. Moreover, Goldfarb's searches are tailored to the particular decade in which the statute was passed.

Writing for the *Atlantic* magazine, commenting on the role of corpus linguistic methods in the *FCC v. AT&T* case, Ben Zimmer, the language columnist for the *Atlantic*, characterized the interpretation of legal texts using empirical, corpus-based data as a "revolution" – a revolution that promises to place "judicial inquiries into language patterns on a firmer, more systematic footing" (Zimmer, 2011).

The Goldfarb's brief in the *FCC v. AT&T* case, and the Supreme Court's apparent reliance on it are important because they demonstrate the Court is receptive to a well-executed presentation of language data in cases about the interpretation of legal texts. Even if the judges did not themselves investigate the interpretive question by directly accessing the corpus, lawyers should take note of the Court's willingness to examine such evidence of meaning.

4.2. The Dictionary Is Not a Fortress

Also in the fall of 2010, my first article entitled *The Dictionary Is Not a Fortress* (Mouritsen, 2010: 1915) was published. The article addressed the question of statutory interpretation from a purely corpus linguistic perspective using data from the Corpus of Contemporary American English ("COCA") and the Corpus of Historical American English ("COHA"). The question addressed in the article was the same question at issue in the *Muscarello* case cited above, namely, the whether the phrase *carries a firearm* ordinarily means to *carry a firearm on your person* or to *carry a firearm in a car*. The defendant in the *Muscarello* case was arrested during a narcotics transaction and received a five-year sentence enhancement for carrying a firearm during the transaction, even though the firearm in question was at all times locked in his glovebox. Writing for the majority, Justice Breyer offered a number of justifications for the conclusion that carry a fire-

arm ordinarily. Justice Breyer argued that because the *conveyance in a vehicle* meaning is the "first definition" in various unabridged English dictionaries, *conveyance* was the term's ordinary meaning (524 U.S. 125 [1998]: 128). This is obviously incorrect as the dictionaries cited by Justice Breyer – the Oxford English Dictionary and the Webster's Third New International Dictionary – rank their definitions historically, oldest to newest. Justice Breyer then refers to *carry*'s etymology arguing that "[t]he ordinary of the word 'carries' explains why the first, or basic, meaning of 'carry' includes conveyance in a vehicle." (id.) Of course, this reasoning is fallacious. Otherwise, December would be the tenth month, not the twelfth (Mouritsen, 2010: 1940).

The article concluded that if the question the ordinary of meaning of *carry a firearm* can be thought of in terms of the frequency of the competing senses, then it is a question that can be addressed with a corpus. The article examined the distribution of senses of *carry* where *carry* is used in the context of *firearm* (or any of the synonyms of firearm – like *rifle, pistol, gun,* etc. – that were attested among the collocates of *carry*). In the COCA, there are six instances of *carry on your person* for every one instance for *carry as conveyance*. This result was amplified when sentences showing only *carry* in the context of *firearm* were examined in the COCA: In that case, there was less than one instance of *carry as conveyance* for every sixty instances of *carry on your person* (Mouritsen, 2010: 1964–1965). These results suggest that the ordinary meaning of *carry a firearm* involves carrying on one's person, contrary to the court's conclusion.

The implications for the *Muscarello* case are profound. While there is only limited data, it is likely that hundreds of people similarly situated to the defendant in *Muscarello* have received the five year sentencing enhancement (Hofer, 2000: 59–62). And the purpose of a judicial opinion is to set forth the Court's justification for its conclusion – a conclusion that in this case upheld a five-year sentencing enhancement. But it is evident from the above that at least some of the justifications given for imposing this sentencing enhancement on the *Muscarello* defendant are not only arbitrary, but deeply erroneous. A prison sentence that is justified, at least in part, on the basis of arbitrary or deeply erroneous reasoning can serve to undermine the public's confidence in the judicial system. This is why predictable and objective approaches interpretation are necessary.

4.3. In re Baby E.Z.

In July of 2011, Justice Thomas R. Lee of the Utah Supreme Court became the first judge to incorporate corpus linguistics into a judicial decision in a case entitled *In re Baby E.Z.* In this case, a biological mother signed a waiver in the State of Virginia relinquishing her parental rights and consenting to an adoption of her child by a Utah couple (*In re Adoption of Baby EZ*, 266 P. 3d 702 [Utah 2011]: 704–705). The child's biological father commenced a custody proceeding in Virginia court, while, a few days later, the adoptive par-

ents commenced an adoption proceeding in Utah. The biological father moved to intervene in the Utah adoption proceeding. The juvenile court denied the request.

On appeal, the biological father raised for the first time a statute called the Parental Kidnapping Prevention Act ("PKPA"), which states:

> "A court of a State shall not exercise jurisdiction in any proceeding for a custody or visitation determination commenced during the pendency of a proceeding in a court of another State […]" (28 U.S.C. § 1738A(g) [2006])

In response to the appeal, the adoptive parents argued that (1) the PKPA applies only to custody proceedings pursuant to a divorce and does not apply to adoption proceedings and that (2) the biological father forfeited his PKPA argument by failing to raise it at the trial court. All five justices agreed that the biological father had forfeited his PKPA argument, but on the question of whether or not the PKPA applies to adoption proceedings, the Court was divided. Writing for the majority, Justice Parrish wrote that "under the plain language of the PKPA, the adoption proceeding below involves a 'custody determination' subject to the PKPA" (266 P. 3d 702 [Utah 2011]: 708).

In a separate concurrence, Justice Lee reached a different conclusion, finding that the PKPA "has no application to adoption proceedings" (id.: 716–724). Justice Lee based this conclusion on a variety of reasons, including the statutory definition, the purpose of the full faith and credit statute upon which the PKPA was premised, the absence of any mention of adoption in the legislative history, and the so-called clear statement rule that requires Utah courts to narrowly construe statutes that implicate traditional state prerogatives like family law.

In addition to these arguments, Justice Lee examined the use of the term *custody* in data from the COCA. In so doing, Justice Lee become the first sitting Judge to rely upon data from a principled linguistic corpus in order to determine the meaning of a word in a statute. Justice Lee first examined the use of *custody* using the KWIC display feature of the corpus (see corpus.byu.edu/coca/?c=coca&q=33387430). "In the context of contemporary usage," he said (266 P. 3d 702 [Utah 2011]),

> "by far the most common family-law sense of the word 'custody' occurs in the setting of a divorce." (724) "This conclusion is based on a review of 500 randomized sample sentences (and the articles or transcripts from which the sentences were drawn) in which the term 'custody' was used in the Corpus of Contemporary American Usage (COCA) *[…]* Of those, 202 uses of the term were found in a criminal law context. One-hundred forty-six explicitly referenced divorce and another seventy-one referenced the actions of child protective services agencies or children placed in foster care. Only twelve sentences out of 500 made any reference to adoption." (724 n. 21)

Justice Lee then proceeded to examine the collocates of the word *custody*. He performed a search similar to that performed by Mr. Goldfarb and determined from that list the likelihood that the word *custody* would occur in the same semantic environment as the words *divorce* and *adoption* (see corpus.byu.edu/coca/?c=coca&q=33387601). "As of this writing," he said, "the COCA reveals 129 co-occurrences of 'custody' with 'divorce,' and only thirteen co-occurrences of 'custody' with 'adoption'" (id.: 724 n.23).

While Justice Lee's opinion garnered some attention and was even heralded as "[a] landmark opinion" (Smith, 2011), Justice Lee's concurrence in the *Baby E.Z.* on the scope of the PKPA did not garner any votes from the other Utah Supreme Court justices. The judges may have had a number of reasons for their skepticism of corpus linguistics, some of which are set forth in the opinion. Certainly, the corpus approach was novel, and novelty is not necessarily an advantage in a tradition-steeped and precedent-based common law system.

Moreover, there was undoubtedly a strong policy argument for applying the PKPA (or a rule like the PKPA) to adoption proceedings. Such a rule would require only that a custody proceeding began in one state would take precedence over any subsequent adoption proceedings in a second state. A legislature could reasonably conclude that such a rule was the best way to serve the interests of the parties and protect the best interests of the child.

But there is no evidence that the legislature ever so concluded:

> "[I]n the hundreds of pages of committee hearings, floor debates, expert testimony, and supporting documentation there is not a single instance in which the word 'adoption' occurs in reference to the PKPA" (266 P. 3d 702 [Utah 2011]: 731 – Lee, J., concurring).

Moreover, the PKPA was passed pursuant to Congress's Full Faith and Credit power, under Article IV, Section 1 of the U.S. Constitution and 28 U.S.C. § 1738, in order to extend "[f]ull faith and credit […] to child custody determinations." (28 U.S.C. § 1738A). Prior to the PKPA, custody determinations were inherently modifiable (266 P. 3d 702 [Utah 2011]: 731 – Lee, J., concurring). One custodial parent could abscond with the child and flee to another state and then get the custody order modified in a new state. The PKPA attempted to put an end to this practice. No such practice could occur in the case of adoption. Adoptions have always been final, unmodifiable judgments, and have always been accorded Full Faith and Credit Status.

Even if the text, structure, and history of the statute make reasonably clear that the PKPA applies only to custody proceedings, what in the end is wrong with a ruling that reaches an admittedly sensible policy outcome, especially one that relies on what some of the judges viewed as a plausible interpretation of the statutory language? This is an important and highly debated question in U.S. jurisprudence. One possible answer, set forth by Professor William N. Eskridge Jr., is "democratic legitimacy":

> "[A]pplying the ordinary meaning of the enacted text of the statute both respects and (possibly) induces accountability of our elected representatives for the statutes they adopt. This value has a formal dimension and a functional one, and they are closely related. Article I, Section 7 of the Constitution provides that congressional bills do not become "law" unless the House of Representatives and the Senate have voted for the same language and have presented that text to the President, whose assent is usually needed unless supermajorities in each chamber override a presidential veto. This constitutional structure, augmented by procedures constitutionally adopted by each chamber, normally assures a great deal of deliberation and compromise for any measure that becomes the law of the land. The normal operation of the legislative process is one where text is supposed to matter a great deal, because the only thing that the House and Senate vote on is statutory text, the best evidence of any rec-

> onciliation of House and Senate versions is the text ultimately adopted, and the only thing presented to the President is the text of the proposed legislation." (Eskridge Jr., 2016: 37)

Judges often state that they must prefer the clear text of a statute over contrary policy preferences (e.g., *Chevron USA Inc. v. Natural Resources Defense Council, Inc.*, 467 U.S. 837 [1984]: 865). Given the importance of such decisions, it seems necessary to have a mechanism to ensure that judges reach predictable and objective conclusions about the meaning of legal texts.

5. Teaching Law and Corpus Linguistics

Though the concurring opinion in *In re Baby E.Z.* did not command the majority of votes in the Utah Supreme Court, the opinion, taken together with the *Atlantic*'s coverage of the corpus linguistics influence in *FCC v. AT&T* and the publication of the *Dictionary Is Not a Fortress* article attracted the attention of then-assistant dean (and current dean) of the J. Reuben Clark Law School at Brigham Young University, Gordon Smith. Dean Smith contacted myself and Justice Lee and proposed the creation of a seminar class on Law and Corpus Linguistics ("LCL") at the BYU Law School. The class seemed like a natural fit for the BYU Law School as the corpora referenced in *In re Baby E.Z.*, the *FCC v. AT&T* amicus brief and related *Atlantic* article, and *The Dictionary Is Not a Fortress* (i.e., the COCA and COHA) were developed at BYU by linguistics professor Mark Davies.

The inaugural course in LCL began in the fall semester of 2013 and we recently completed its fourth year in the fall semester of 2016.[10] As a seminar course, students attend a weekly lecture and are expected by the end of the semester to produce original research in the field of LCL. The lectures cover a number of potential applications for linguistic corpora in the law, including the use of corpora in the interpretation of contemporary legal texts, such as statutes, contracts, and agency rules, and the use of corpora in the interpretation of historical texts, including the U.S. Constitution and its various amendments. The lectures also address additional potential applications of corpus linguistics in the fields such as trademark, contract, and agency law. The course also addresses areas in which the use of linguistic corpora are already well-established, including areas such as political discourse and forensic linguistics. The course is taught with a strong emphasis on applied corpus linguistics. Questions of legal interpretation are discussed in class and students are expected to use linguistic corpora in class to address these problems. By the end of each semester students are expected to have prepared a paper addressing at least one legal or interpretive issue through the use of linguistic corpora (e.g., Ortner, 2016: 101).

[10] I teach the course together with Justice Lee and Dean Smith.

The purpose of this course is to teach a younger generation of lawyers to look at interpretative problems in a new way. As we saw with some early responses to corpus linguistic approaches to corpus-based interpretation were met with skepticism, in part because they were encountered by judges and lawyers immersed in a tradition-steeped and precedent-based common law system that tends to look to the past for answers and not to the future. While some of the courses students continue to work to publish original corpus-based research, each leaves the class with an understanding of new ways to look at old questions of interpretation.

6. *State v. Rasabout* and the Emergence of Law and Corpus Linguistics

During the follow up period after the *In re Baby E.Z.* opinion, there was very little mention of LCL in judicial opinions and academic writing in the United States.[11] Then, in 2015, the Utah Supreme Court issued its opinion in *State v. Rasabout* (2015 UT 72, 356 P.3d 1258).

In *Rasabout*, the Utah Supreme Court was called upon to determine the unit of prosecution for a statutory prohibition against the "discharge of a firearm." Utah Code § 76-10-508. That is, the defendant in the *Rasabout* case had fired his gun twelve times, and the question before the court was whether these twelve shots constituted a single "discharge" or twelve separate "discharge[s]" for which the defendant could be prosecuted (id.: 2–3). In a lengthy concurring opinion, Justice Lee again uses corpus linguistics to address the linguistic uncertainty in the *Rasabout* case (id.: 88–93). He concludes that

> "[b]y examining the instances of *discharge* in connection with these nearby nouns, I confirmed that the single shot sense of this verb is overwhelmingly the ordinary sense of the term in this context." (id.)

More importantly, Justice Lee spends a considerable portion of his lengthy concurrence defending the use of corpus linguistics against the allegation that corpus linguistics inquiries are barred by ethics rules against judges in an adversarial system from investigating facts and that corpus linguistics is "scientific field of study" best left to the experts (id.: 101).

Justice Lee responded that evidentiary rules prevent judges in an adversarial system from investigating adjudicative facts, but not legislative ones – *i.e.*, facts that go to the meaning and purpose of the law (id.: 105). Judges are expressly permitted to research

[11] There were exceptions. Rather than engage in a full-fledged corpus linguistics approach using a principled corpus like the COCA, Justice Lee relied on a quasi-corpus search of a Google News archive to address the meaning of "out of state" in his majority opinion in the case of *State v. Canton* (2013 UT 44: 26–27 – 308 P.3d 517). Also, during the period, I published my second LCL paper (Mouritsen, 2011: 202) addressing the meaning of "enterprise" in the Racketeer Influenced and Corrupt Organizations Act ("RICO"), 18 U.S.C. §§ 1961–1968.

so-called legislative facts, and the meaning, purpose, and interpretation of the text of the law have always been questions for the judge to resolve (id.). With respect to whether or not corpus linguistics is properly the domain of experts, Justice Lee responds:

> "We judges are experts on one thing – interpreting the law. And the fact that that enterprise may implicate disciplines or fields of study on which we lack expertise is no reason to raise the white flag. It is reason to summon all our faculties as best we can, and to overcome any weaknesses we may possess. This is not a matter of dreaming up 'interesting research projects.' It is a matter of doing our job" (id.: 108)

Like the *Muscarello* case, the opinion in *Rasabout* will have a dramatic effect not only on the defendant in that case, but on all others for whom the unit of prosecution may now be amplified. Where such important liberty interests are dependent on the interpretation of a single text, it is vital that the interpretation of that text be conducted in as predictable and objective manner as possible. Arbitrary and institution based reasoning about ordinary meaning should not be the exclusive basis for significantly enhancing an individual's exposure to criminal liability. In this respect, a corpus-based approach to interpretation may be one way to check a judge's intuition and prevent arbitrary reasoning about the meaning of a text.

The debate about LCL in the competing opinions in the *Rasabout* case attracted significant attention in the legal academy in the U.S. The case was discussed in the Harvard Law Review (Note, 2016: 1468), and discussed on a number of prominent legal blogs, including the Washington Post's Volokh Conspiracy (Volokh, 2015), the National Review's Bench Memos (Whelan, 2015), and The Conglomerate (Smith, 2016). Shortly after the opinion was issued, essays debating the use of historical corpora to interpret the U.S. Constitution were published in the Yale Law Journal Forum (Phillips, Ortner & Lee, 2016: 21; Solan, 2016: 57). In addition, a recent treatise by a leading figure in statutory interpretation, Professor William N. Eskridge Jr., addressed the issue of corpus-based interpretation (Eskridge Jr., 2016: 45–47).

The following spring, the BYU Law School, together with the Center for the Constitution at the Georgetown University Law Center, hosted the first ever U.S. academic conference on LCL.[12] Professor Larry Solum, the head of Georgetown's Center for the Constitution, said of the conference that it was

> "an important and path breaking event – the first in my knowledge to undertake a systematic exploration of corpus linguistics and the interpretation of legal texts." (Solum, 2016)

[12] Corpus Linguistics Conference, BYU Law School (May 3, 2016), see http://www.law2.byu.edu/news2/corpus-linguistics-conference. Previously, international conferences related to LCL have been hosted by the Computer Assisted Legal Linguistics (CAL²) International Research Group: "Legal Corpus Pragmatics: Corpus-Based Approaches to Legal Semantics" at the Freiburg Institute for Advanced Studies ("FRIAS") at the Albert-Ludwigs-University (Freiburg, Germany), April 25–27, 2013; The Fabric of Language and Law: Discovering Patterns Through Legal Corpus Linguistics (Heidelberg, Germany), March 18–19, 2016.

The conference brought together academics from the fields of both law and linguistics with the aim of encouraging participants to conduct original research. Many of the participants in this first conference would present their original research nearly a year later at a second LCL conference hosted again at BYU.[13]

Not long after the first BYU LCL conference, the Michigan Supreme Court adopted a corpus-based approach to statutory interpretation, relying on the data from the COCA to interpret a statute proscribing the use of "information" obtained from police officers during internal investigations in subsequent criminal proceedings (*People v. Harris*, 885 N.W.2d 832 [2016]). The court stated:

> "Keeping in mind that we must interpret the word 'information' as used in the [statute] 'according to the common and approved usage of the language,' we apply a tool that can aid in the discovery of 'how particular words or phrases are actually used in written or spoken English. The Corpus of Contemporary American English (COCA) allows users to 'analyze[] ordinary meaning through a method that is quantifiable and verifiable.'" (838–839)

Both the majority and the dissent relied on corpus data,[14] and Justice Zahra, author of the majority opinion, would go on to lecture about the benefits of a corpus-based interpretive method before the Michigan Bar (see Levy, 2016; Thomas, 2016: 60).

The Michigan Supreme Court's decision in *People v. Harris* is remarkable because both the majority and dissent relied on corpus data, but reached opposite conclusions. If corpus-based interpretation is ostensibly predictable and objective, how did these judges reach separate opinions after examining the same data? The answer is that the judges drew the same conclusions observations from the data, but reached different conclusions about what constitutes "ordinary meaning." The majority stated:

> "Empirical data from the COCA, however, demonstrates [… that in] common usage, 'information' is regularly used *in conjunction with adjectives suggesting it may be both true and false*. This strongly suggests that the unmodified word 'information,' *can* describe either true or false statements." (885 N.W.2d 832 [2016]: 839)

To this the dissent responded that

> "99.44% of the time 'information' in the COCA is unmodified by any of these adjectives related to veracity […] And where 'information' is unmodified by one of these adjectives, I believe it is overwhelmingly used to refer to truthful information. See, e.g., the utterly ordinary, commonplace, and pedestrian usages of "information" set forth in the COCA." (id.: 850 n.14 – Markman, J., dissenting)

That is, the majority found that "information" is sometimes modified by adjectives related to veracity and at least sometimes can mean either "true" or "false" information. The dissent observed that in the overwhelming majority of cases, information is un-

[13] BYU Law & Corpus Linguistics (February 3, 2017), at lawcorpus.byu.edu. Papers by Solum, forthcoming 2017; Gries & Slocum, forthcoming 2017; Solan & Gales, forthcoming 2017; Hamann & Vogel, forthcoming 2017; Mascott, forthcoming 2017; Goldfarb, forthcoming 2017; Strang, forthcoming 2017; Phillips & Egbert, forthcoming 2017.

[14] See 885 N.W.2d 832 [2016]: 850 n.14 (Markman, J., dissenting): "the Corpus of Contemporary American English (COCA), a truly remarkable and comprehensive source of ordinary English language usage".

modified and in those cases almost always means "truthful information." At bottom, the *Harris* case may represent a disagreement, not about the meaning of "information," but about the meaning of ordinary meaning.

Finally, after the publication of the decision in *People v. Harris*, a majority of the Utah Supreme Court signaled that it would welcome corpus-based briefing: "All agree that our analysis of [corpus linguistics] (or any other issue) will be enhanced by adversary briefing." (*Craig v. Provo City*, 2016 UT 40: 26 n.3)

7. Challenges and the Future of Law and Corpus Linguistics

In order for corpus linguistics to be woven into the fabric of legal interpretation, its proponents must first anticipate some likely criticisms. Among these is the question of whether a corpus consisting of non-legal texts should be used as a basis for resolving normative questions in legal texts that are, presumably, written is specialized, legal language.

This concern is understandable, but in there is a long tradition of resolving disputes about the meaning of legal texts with reference to language used by the community at large, rather than according to the specialized, legal conventions. This tradition was expressed by United States Supreme Court Justice Oliver Wendell Holmes, in the case of *McBoyle v. United States*, in which Justice Holmes stated:

> "Although it is not likely that a criminal will carefully consider the text of the law before he murders or steals, *it is reasonable that a fair warning should be given to the world in language that the common world will understand*, of what the law intends to do if a certain line is passed. To make the warning fair, so far as possible the line should be clear." (*McBoyle v. U.S.*, 283 U.S. 25 [1931]: 27 – emphasis added)

There are good reasons that U.S. courts attempt to apply the ordinary meaning (as opposed to a specialized, legal meaning) when interpreting generally applicable federal statutes. Professor William Eskridge Jr. has stated:

> "There are excellent reasons for the primacy of the ordinary meaning rule. To begin with, ordinary meaning matches up well with our understanding of what the *rule of law* entails. A polity governed by the rule of law aspires to have legal directives that are known to the citizenry, that are predictable in their application, and that officials can neutrally and consistently apply based upon objective criteria […] For this reason, there is perhaps no more important role for legislators and administrators than to generate well-understood rules that guide people's conduct into productive channels, and no more important role for judges than to enforce those rules through a method that is objective, general, and predictable." (Eskridge Jr., 2016: 35)

Professor Eskridge continues, quoting Justice Holmes, to observe that "the primary task for the statutory interpreter is to determine 'what [the statutory] words would mean in the mouth of an ordinary speaker of English, using them in the circumstances in which they were used'," and adds: "This foundational rule for America's republic of

statutes is a strong presumption that We the People as well as government officials ought to read statutes in accord with the ordinary meaning their words and phrases would have for the typical English-speaking citizen" (Eskridge Jr., 2016: 41). Moreover, legislative drafters compose new statutes with this "foundational rule" in mind (Eskridge Jr., 2016: 41, citing Nourse & Schacter, 2002: 594–597).

Because U.S. judges and lawyers have a long tradition of interpreting legal texts according to their ordinary meaning, and because legislative drafters create new statutes with this rule in mind, access to linguistic corpora may assist judges in discovering the linguistic norms and conventions of the community at large.

This is not to suggest that the ordinary meaning of a text should always prevail. Numerous cases recognize that

> "where Congress borrows terms of art in which are accumulated the legal tradition and meaning of centuries of practice, it presumably knows and adopts the cluster of ideas that were attached to each borrowed word in the body of learning from which it was taken and the meaning its use will convey to the judicial mind unless otherwise instructed." (Eskridge Jr., 2016: 60, quoting *Morissette v. U.S.*, 342 U.S. 246 [1952]: 253, and other sources in n.63)

One could argue that a corpus of non-legal texts would be unhelpful. However, U.S. courts have no systematic way for identifying if and when specialized legal meaning should attach to a given utterance. Here, comparative legal and non-legal corpora might help render the identification and interpretation of legal terms of art more systematic.

There are other challenges. Judges are specialists in the law, but generalists, at best, when it comes to linguistics. As Judge Frank Easterbrook has observed:

> "Judges are overburdened generalists, not philosophers or social scientists. Methods of interpretation that would be good for experts are not suitable for generalists." (Easterbrook, 1994: 67)

It is appropriate to ask whether judges can, and should, develop sufficient expertise to employ and understand corpus methods in interpreting statutes. However, this create seems to miss an important point. Though judges are generalists with respect to many of the issues that come before them, they are expected to be specialists, even experts, with respect to interpretive tasks. If traditional methods of interpretation can be shown to be inadequate, judges cannot shy away from the task of learning new methods simply by hiding under the title of generalists. Judges are specialists when it comes to interpretation and can be expected to learn effective methods for reaching predictable and objective outcomes to interpretive problems.

Finally, there is a potential concern that judges in an adversarial system should not be conducting independent research about the meaning of a statute, but should instead rely only on arguments and interpretations presented by counsel. But as Justice Lee noted in the *Rasabout* case above, judges while judges in an adversarial system are not permitted to independently investigate facts, the interpretation of the meaning of a legal text has always been legal question and the sole responsibility of judges. Just as

judges had to learn to rely on legal software to research case law and precedent, judges may one day turn to linguistic corpora to address questions or ordinary meaning.

Writing in 2004, Professor Lawrence Solan made the following prediction about the future of LCL:

> "Over the past decade, a great deal of work has been published in the growing field of corpus linguistics [...] Access to computers now makes it relatively simple to see how words are used in commerce and in common parlance. This allows judges to easily become their own lexicographers. If they perform that task seriously, they stand to learn more about how words are ordinarily used, than by today's method of fighting over which dictionary is the most authoritative" (Solan, 2005: 2059–2060).

Professor Solan's prediction that judges might one day "become their own lexicographers" has begun to take shape. Judges are already turning to linguistic corpora to learn more about language usage and to better and more objectively perform the task of interpreting legal texts. But if this trend is going to continue, then legal theory must keep pace with advances in our understanding of human language and advances in language technology. We must begin to fill in gaps in interpretative theory. Corpus linguistics can provide a sample of the speech of a given speech community at a given point in time. But what is the appropriate speech community to consider when interpreting a statute – the speech of the trained legal professionals who write the laws, or the speech of the ordinary citizen that is subject to the laws in question? Should the interpretation of a contract take into account the relative sophistication of each party, and should differences in education, or even geographic origin of the parties be taken into account? If so, how can these factors be empirically and objectively accounted for in corpus design? Finally, what is the proper role of judges, experts, and the parties when corpus data is used in an adversarial setting?

Legal scholars are only now beginning to answer these questions. But the promise of the LCL movement is that when such answers come, they will be grounded not merely on impressionistic arguments, but instead will be grounded in empirical data gathered through experiments that are both replicable and falsifiable and therefore satisfy the highest values of the scientific method.

References

Bintliff, Barbara (1996). From Creativity to Computerese: Thinking Like a Lawyer in the Computer Age. *Law Library Journal, 88*, 338–351.

Brudney, James J. & Baum, Lawrence (2013). Oasis or Mirage: The Supreme Court's Thirst for Dictionaries in the Rehnquist and Roberts Eras. *William & Mary Law Review, 55*, 483–580. Available at wmlawreview.org/oasis-or-mirage.

Dickerson, F. Reed (1961). The Electronic Searching of Law. *American Bar Association Journal, 47*, 902–908. Available at repository.law.indiana.edu/facpub/1503.

Dickerson, Reed (1983). Statutory Interpretation: Dipping Into Legislative History. *Hofstra Law Review, 11*, 1125–1162. Available at hofstralawreview.org/archive/volume-11-issue-4-summer-1983.

Easterbrook, Frank H. (1994). Text History, and Structure in Statutory Interpretation. *Harvard Journal of Law and Public Policy, 17*, 61–70. Available at chicagounbound.uchicago.edu/journal_articles/1170.

Eskridge Jr., William (2016). Interpreting Law: A Primer on How to Read Statutes and the Constitution. St. Paul, MN: Foundation Press.

Goldfarb, Neal (forthcoming 2017). Words, Meanings, Corpora: A Lawyer's introduction to Meaning in the Framework of Corpus Linguistics. *Brigham Young University Law Review*.

Gries, Stefan Th. & Slocum, Brian (forthcoming 2017). Ordinary Meaning and Corpus Linguistics. *Brigham Young University Law Review*.

Hamann, Hanjo & Vogel, Friedemann (forthcoming 2017). Evidence-Based Jurisprudence Meet s Legal Linguistics—Unlikely Blends Made In Germany. *Brigham Young University Law Review*.

Hart Jr., Henry M. & Sacks, Albert M. (1994). *The Legal Process: Basic Problems in Making and Application of Law* (Eskridge, Jr. & Frickey eds.). Westbury, NY: Foundation Press.

Hietala Jr., James R. (2014). Linguistic Key Words in E-Discovery. *American Journal of Trial Advocacy, 37*, 603–620.

Hofer, Paul J. (2000). Federal Sentencing for Violent and Drug Trafficking Crimes Involving Firearms: Recent Changes and Prospects for Improvement. *American Criminal Law Review, 37*, 41–74.

Kilgarriff, Adam (2007). Googleology Is Bad Science. *Computational Linguistics, 33*(1), 147–151. DOI: 10.1162/coli.2007.33.1.147.

Kredens, Krzysztof & Coulthard, Malcolm (2012). Corpus Linguistics in Authorship Identification. In Solan & Tiersma (Eds.), *The Oxford Handbook of Language and Law* (pp. 489–510). Oxford (UK): Oxford University Press.

Lee, Thomas R. & Mouritsen, Stephen C. (forthcoming 2017). Judging Ordinary Meaning. *Yale Law Journal, 126*.

Leonard, Robert A. (2008). Declaration in Opposition to Microsoft Corp.'s Motion for Summary Judgment, In the Matter of Application Serial No. 77/525,433.

Levi, Judith (2008). Expert Declaration in Support of Whirlpool Corporation's Memorandum of Law Opposing LG's Motion for Preliminary Injunction. *LG Electronics U.S.A. v. Whirlpool Corp.*, No. 08-C-2008 WL 670474 (N.D. Ill.)

Levy, Douglas (2016). Zahra Instructs Lawyers on Corpus Linguistics. *Michigan Lawyers Weekly*, 5 Oct.

Lien, Molly Warner (1998). Technocentrism and the Soul of the Common Law Lawyer. *American University Law Review, 48*, 85–86. Available at aulawreview.org/pdfs/48/48-1/lien.pdf.

Mascott, Jennifer L. (forthcoming 2017). The Dictionary as a Specialized Corpus. *Brigham Young University Law Review*.

McEnery, Tony & Wilson, Andrew (2001). *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh (UK): Edinburgh University Press.

Melton, Jessica S. & Bensing, Robert C. (1961). Searching Legal Literature Electronically: Results of a Test Program. *Minnesota Law Review, 45*, 229–248.

Mouritsen, Stephen (2010). The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning. *Brigham Young University Law Review*, 1915–1979. Available at digitalcommons.law.byu.edu/lawreview/vol2010/iss5/10.

Mouritsen, Stephen (2011). Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning. *Columbia Science and Technology Law Review, 13*, 156–205. Available at stlr.org/cite.cgi?volume=13&article=4.

Note (1967). The Use of Data Processing in Legal Research. *Michigan Law Review, 65*, 987–994. DOI: dx.doi.org/10.2307/1287094.

Note (1993–1994). Looking It Up: Dictionaries and Statutory Interpretation. *Harvard Law Review, 107*, 1437–1454. DOI: 10.2307/1341851.

Note (2016). Statutory Interpretation—Interpretative Tools—Utah Supreme Court Debates Judicial Use of Corpus Linguistics—*State v. Rasabout*, 356 P.3d 1258 (Utah 2015). *Harvard Law Review*, *129*, 1468–1475. Available at harvardlawreview.org/2016/03/state-v-rasabout.

Nourse, Victoria F. & Schacter, Jane S. (2002). The Politics of Legislative Drafting: A Congressional Case Study. *New York University Law Review*, *77*, 575–624. Available at nyulawreview.org/issues/volume-77-number-3.

O'Keeffe, Anne & McCarthy, Michael (2010). *The Routledge Handbook of Corpus Linguistics*. Hoboken, NJ: Taylor & Francis.

Ortner, Daniel (2016). The Merciful Corpus: The Rule of Lenity, Ambiguity and Corpus Linguistics. *Boston University Public Interest Law Journal*, *25*, 101–142.

Phillips, James C., Ortner, Daniel M. & Lee, Thomas R. (2016). Corpus Linguistics & Original Public Meaning: A New Tool To Make Originalism More Empirical. *Yale Law Journal Forum*, *126*, 21–32. Available at yalelawjournal.org/forum/corpus-linguistics-original-public-meaning.

Phillips, James C. & Egbert, Jesse (forthcoming 2017). Improving Corpus Design and Corpus-Based Analysis for Linguists and Lawyers: Principles and Practices from Survey and Content-Analysis Methodologies. *Brigham Young University Law Review*.

Posner, Richard A. (2013). *Reflections on Judging*. Cambridge, MA: Harvard University Press.

Smith, Gordon (2011). A Landmark Opinion: Corpus Linguistics in the Courts. *The Conglomerate*, 19 Jul. Available at theconglomerate.org/2011/07/a-landmark-opinion-corpus-linguistics-in-the-courts.html.

Smith, Gordon (2016). Michigan Supreme Court Embraces Corpus Linguistics. *The Conglomerate*, 28 Jun. Available at theconglomerate.org/corpus-linguistics.

Solan, Lawrence M. (2005). The New Textualist's New Text. *Loyola of Los Angeles Law Review*, *38*, 2027–2062. Available at digitalcommons.lmu.edu/llr/vol38/iss5/5.

Solan, Lawrence M. (2016). Can Corpus Linguistics Help Make Originalism Scientific? *Yale Law Journal Forum*, *126*, 57–64. Available at yalelawjournal.org/forum/can-corpus-linguistics-help-make-originalism-scientific.

Solan, Lawrence M. & Gales, Tammy (forthcoming 2017). Corpus Linguistics as a Tool in Legal Interpretation. *Brigham Young University Law Review*.

Solum, Lawrence B. (2016). Conference Hopping: BYU, Melbourne, Monash, and Chicago. *Legal Theory Blog*, 1 May. Available at lsolum.typepad.com/legaltheory/2016/05/conference-hopping-byu-melbourne-monash-and-chicgo.html.

Solum, Lawrence B. (forthcoming 2017). Originalist Methodology and Corpus Linguistics. *Brigham Young University Law Review*.

Strang, Lee J. (forthcoming 2017). The Original Meaning of "Religion" in the First Amendment: A Test Case for Originalism's Utilization of Corpus Linguistics. *Brigham Young University Law Review*.

Sunstein, Cass R. (1997). Behavioral Analysis of Law. *The University of Chicago Law Review*, *64*, 1175–1196. Available at chicagounbound.uchicago.edu/journal_articles/8314.

Thomas, Virginia C. (2016). Of Plain English and Plain Meaning. *Michigan Bar Journal*, *95*, 60–61. Available at michbar.org/journal/home/VolumeId=195.

Thornburg, Elizabeth G. (2008). The Curious Appellate Judge: Ethical Limits on Independent Research. *The Review of Litigation*, *28*, 131–202. Available at ssrn.com/abstract=1267684.

Thumma, Samuel A. & Kirchmeier, Jeffrey L. (1999). The Lexicon Has Become a Fortress: The United States Supreme Court's Use of Dictionaries. *Buffalo Law Review*, *47*, 227–561. Available at ssrn.com/abstract=920511.

Thumma, Samuel A. & Kirchmeier, Jeffrey L. (2010). Scaling the Lexicon Fortress: The United States Supreme Court's Use of Dictionaries in the Twenty-First Century. *Marquette Law Review*, *94*, 77–262. Available at ssrn.com/abstract=1832926.

Véronis, Jean (1998). A Study of Polysemy Judgements and Inter-Annotator Agreement. *Programme and Advanced Papers of the Senseval Workshop, Herstmonceux*. Available at pdfs.semanticscholar.org/ac52/5c6ed4034565642l5bf1f32924032d68f427.pdf.

Volokh, Eugene (2015). Judges and 'corpus linguistics'. *The Volokh Conspiracy*, 17 Aug. Available at washingtonpost.com/news/volokh-conspiracy/wp/2015/08/17/judges.

West, John B. (1909). Multiplicity of Reports. *Law Library Journal, 2*, 4–7.

Whelan, Ed (2015). Corpus Linguistics as Interpretive Tool. *Bench Memos*, 19 Aug. Available at nationalreview.com/bench-memos/422755/corpus-linguistics-interpretive-tool-ed-whelan.

Zimmer, Ben (2011). The Corpus in the Court: 'Like Lexis on Steroids'. *The Atlantic*, 4 Mar. Available at theatlantic.com/national/archive/2011/03/the-corpus-in-the-court-like-lexis-on-steroids/72054.

International Journal of
Language & Law

Marín, JLL 6 (2017): 18–45
www.languageandlaw.de

Legalese as Seen Through the Lens of Corpus Linguistics

— An Introduction to Software Tools for Terminological Analysis

*María José Marín**

Abstract
In spite of the plethora of possibilities offered by Corpus Linguistics to the study of legal English, the research devoted to the study of this English variety based on this discipline is not as fruitful as that dedicated to other branches of ESP. The present research could be regarded as an introduction into major issues related to the design and compilation of a legal corpus such as the application of appropriate sampling strategies to ensure its representative value. This study also examines the implementation of Automatic Term Recognition (ATR) methods for the analysis of legal terminology and the automatic deployment of collocate networks. The first section explores such a controversial issue as establishing the ideal size for a specialised corpus applying the type/term ratio to a corpus of judicial decisions, the *BLaRC*, used as reference. In section 3, the assessment of different Automatic Term Recognition (ATR) methods is described. Out of five different methods, Drouin's (2003) *TermoStat* is found and recommended as the most efficient one in legal term mining. Finally, sections 4 and 5 demonstrate the practicality of collocate networks (Williams, 1998; 2001) in their capacity to reveal lexico-grammatical patterns which provide plenty of information for the study of legal text. A case study of the sub-technical legal term *party* using *Lancsbox* – designed by Brezina, McEnery & Wattam (2015) – is presented in section 5.2, where its general and specialised contexts are examined. Such scrutiny brings to the foreground interesting data such as the relevance of marriages of convenience in a collection of judicial decisions.

Keywords
Legal English, Corpus Linguistics, Terminology, Automatic Term Recognition, Collocate Networks, Lancsbox

Submitted: 27 October 2016, accepted: 31 July 2017, published online: 13 August 2017

* University of Murcia, Spain, mariajose.marin1@um.es.

1. Introduction

As commonly agreed by scholars, legal English (also known as *legalese*) is a peculiarly obscure and convoluted variety of English. David Mellinkoff, one of the first scholars devoted to the study of *legalese*, affirms that "the language of the law has a strong tendency to be: wordy; unclear; pompous [and] dull" (Mellinkoff, 1963: 63). The presence of Latin borrowings and Old French phrases, synonyms, archaisms and redundancy, as well as the widespread use of "common words with uncommon meanings" (Mellinkoff, 1963: 11) characterise its lexicon.

Traditionally, most of the work devoted to the description of legal English features (Mellinkoff, 1963; Alcaraz, 1994; Tiersma, 1999; Borja, 2000) has been either based on the authors' knowledge and intuitions on the subject or on relatively reduced language samples. These studies have often presented a top-down characterisation of the major traits of this ESP variety, following a deductive approach whereby the rule usually precedes the actual description of the examples provided. Nevertheless, there is a growing tendency towards corpus-based and corpus-driven[1] descriptions of *legalese* which provide a bottom-up characterisation of this ESP branch (Marín & Rea Rizzo, 2012; Biel & Engberg, 2013; Goźdź-Roszkowski & Pontrandolfo, 2014; Breeze, 2015).

Scholars have profusely discussed the advantages and disadvantages of employing language corpora as a source of information for linguistic analysis (Sinclair, 1991; McEnery & Wilson, 1996; Dudley-Evans & St. John, 1998; Kennedy, 1998; McEnery, Xiao & Tono, 2006; Tognini-Bonelli, 2001; Gries & Wulff, 2010). The Chomskyan distinction between competence and performance stands at the very basis of the earliest criticism against this discipline, which can be traced back to the 50s and 60s. Following Chomsky (1965), intuitive examples, as traditionally formulated by linguists, reflect linguistic competence as they arise from our tacit knowledge of the system and should serve as dependable references to base language theory upon. Conversely, those examples taken from corpora reflect performance, which usually mirrors competence poorly. As Chomsky puts it,

> "the problem for the linguist (…) is to determine from the data of performance the underlying system of rules that have been mastered by the speaker-hearer and that he puts to use in actual performance" (1965: 4).

Along these lines, some authors supporting this attitude have often deemed corpus samples skewed, frequently leading the linguist to erroneous generalisations on the language and offering "truncated concordance lines [which] are examined atomistically" (Flowerdew, 2009: 395). However, as Widdowson (2000) acknowledges, neither purely intuitive approaches to language description nor those based uniquely on

[1] In corpus-based linguistic studies a query is formulated in advance so as to find evidence in a corpus, whereas corpus-driven analyses base their conclusions solely on linguistic findings obtained from corpora and adopt an inductive approach to language description.

Corpus Linguistics are complete without each other. As a matter of fact, what the latter can do

> "is reveal the properties of text, and that is impressive enough. But it is necessarily only a partial account of real language. For there are certain aspects of linguistic reality that it cannot reveal at all. In this respect, the linguistics of the attested is just as partial as the linguistics of the possible" (Widdowson, 2000: 7).

In spite of earlier criticism and due to the fast growth of corpora and processing software nowadays, researchers can rapidly access and analyse large amounts of data that could have not even been thought of in the 50s and 60s. Tools like *Sketch Engine* (Kilgarriff et al., 2014) allow the user to search keywords, collocate patterns (sketches) and concordance lines employing as reference gigantic corpora like *enTenTen12*, of 12 billion words. Such plethora of data grants the reliability of the conclusions drawn from the observation of the language samples thus obtained, although the degree to which corpus data should be employed as the only source to base language description upon still remains an open question. In our view, intuition should go hand in hand with data collection, as remarked by Partington (1998), and aid the researcher, for instance, to discard ungrammatical examples. Similarly, the direct observation of the data can also contribute to the confirmation of hypotheses or *a priori* formulated theories and call our attention to new aspects of the language that could not be detected otherwise.

The applications offered by Corpus Linguistics to the study of general and specific languages are manifold, allowing for a descriptive approach to real language usage and also for the processing of large amounts of text. Nevertheless, the techniques and tools available may not always be well-known or easy to handle for non-specialists in the field such as law practitioners or linguists not accustomed to using corpora as part of their research methodology. This study was thus conceived as an introduction into this linguistic discipline for the analysis of legal English, especially aimed at those researchers unfamiliar with the wide array of corpus analysis tools available and the number of possibilities they offer.

Section 2 of this paper offers a general overview on such fundamental questions related to corpus design as how to determine the ideal size of a corpus or how to structure it. Additionally, section 3 presents a reflection on the usefulness of automatic term recognition tools by assessing their efficiency in legal term extraction. In section 4, the work by Williams (1998; 2001) and Brezina, McEnery & Wattam (2015) on collocational networks is presented. The article concludes with a case study of the term *party* in the general and the specialised fields using the software package *Lancsbox* (Brezina, McEnery & Wattam, 2015), which enables the user to obtain the lexical network of a given word/term and extend its context of usage up the seventh collocational level.

The three research questions (RQs) which motivated this study are the following:

RQ1: What key issues must be considered in the design and compilation of a legal corpus? How can they be tackled?

RQ2: How can automatic term recognition methods contribute to the study of legal texts? Can we trust these methods as dependable tools to rely on?

RQ3: How can collocation patterns add to the study of legal texts? Are there any automatic tools which facilitate such task?

2. Corpus description and justification

Answering the first research question on the most relevant issues to be considered in the design and compilation of a specialised corpus and how to tackle them is not an easy task. There seems to be general agreement on the importance of applying the appropriate sampling strategies in the selection of texts, since using a reliable method in corpus design is fundamental for the results obtained from its analysis to be representative of a given language variety. Biber (1993; 1998), McEnery & Wilson (2001), Sinclair (2005), McEnery, Xiao & Tono (2006), Tognini-Bonelli (2001) or Gries & Wulff (2010), to name but a few, provide a detailed insight into such and other issues, which are seminal in Corpus Linguistics. Following these authors, there are questions such as establishing the word targets or considering the communicative relevance of the text types included in a corpus that must be carefully tackled in its design and compilation.

This section presents a discussion on some of these issues[2] and the decision-making process in the design of the *British Law Report Corpus* (*BLaRC* henceforth), the legal text collection employed in this research.

2.1. Communicative relevance of law reports in common law legal systems

The *BLaRC*,[3] an 8.5 million word legal English corpus containing 1,228 legal texts, is a collection of British law reports issued by British courts between the years 2008 and 2010. Law reports are collections of judicial decisions or judgments which stand at the very core of common law systems and act as the main source of law followed by statutes and equity, hence their relevance within the British system. Following Sinclair, "the contents of the corpus should be selected […] according to their communicative function in the community in which they arise" (in Wynne, 2005: 5), a statement which insists on the aptness of this genre for the compilation of a legal corpus.

[2] See Marín & Rea Rizzo (2012) for further details.

[3] The corpus is freely available online at http://lextutor.ca/conc/eng and http://flax.nzdl.org/greenstone3/flax .

The United Kingdom belongs to the realm of common law, where judicial decisions are based on previous cases always abiding by the doctrine of *stare decisis* (to stand by what has previously been decided) or principle of biding precedent. The decisions made by a higher court should act as binding precedent as long as they are related to the case in question in their essence. Determining what the essence of a given case is, that is, establishing the *ratio decidendi*, is part of the judge's role. "Cases must be decided the same way when their material facts are the same, [...] but the legally material facts may recur and it is with these that the doctrine is concerned", according to Williams (in Bhatia, 1993: 128). Nevertheless, judges are also subject to statutory principles, which must be interpreted whenever applicable and also act as a source of law. Statutory law has gained relevance as a major legal source in the UK in the last 150 years (Geary & Morrison, 2012; Orts, 2006), even so, law reports still stand at the very basis of the legal system and legal practitioners must know them well.

Actually, law reports must be cited and act as one of the essential elements which lawyers build their arguments upon and judges base their decisions on. This is why, in the UK, they are made public through different institutions, both public and private, i.e., the Incorporated Council of Law Reports of England and Wales (ICLR) or publishing houses like Butterworth or Lloyds. Due to the widespread use of information technologies, there is a tendency towards digitalising these texts and storing them in online databases. The British and Irish Legal Information Institute (bailii.org) offers an open-access online database where the judicial decisions made at British courts (as well as many other documents from various sources) can be consulted and downloaded.

As regards the generic classification of law reports, it varies depending on the perspective adopted for their analysis. Law reports may appear in generic classifications as part of the oral mode (Danet, 1980), within the category "recording and law making" (Maley, 1994) or as public unenacted law (Orts, 2009), amongst others.

Another relevant communicative function of law reports, as highlighted by Bhatia (1993) and Nesi & Gardner (2012), is the role they play within Higher Education. Becoming a solicitor or a barrister in the UK requires passing a hard process of accreditation which law faculties prepare students for. Amongst many other requirements, the suitors must be able to write case reports, thus having to apply and cite law reports as the major source to base their arguments on. Writing case reports is not only part of their training but also of their professional activity although only barristers can "be called to the bar", that is, argue a case in court on behalf of their clients.

Finally, law reports are rather comprehensive texts since they not only cover all the branches of law, but also present full sections of other legal texts such statutes, wills, contracts, deeds and the like. Nesi & Gardner (2012: 177) provide a description of the macrostructure of law reports which follow four principal stages:

i) case identification;
ii) case facts;

iii) arguing of the case (case history, presentation of arguments, *ratio decidendi*), and
iv) judgement.

Citing sections of statutes or the contents of some other private documents is a usual procedure when arguing a case, hence the relevance of this legal genre not only from a legal but also from a linguistic point of view if a terminological study (like the one presented below) is to be carried out.

2.2. Corpus size and representativeness: establishing the word target

Representativeness is vital in corpus design. Douglas Biber (1993) – a fundamental reference in this field – refers to the crucial role performed by corpus sampling strategies, which may be decisive to determine whether a corpus is representative of the variety of the language it aims at covering or simply an illustrative sample of it with no predictive value. Biber insists on the transcendence of this issue owing to the fact that "representativeness refers to the extent to which a sample includes the full range of variability in a population" (Biber, 1993: 246).

Therefore, the concept *representative*, as defined by Biber, points at two major questions, on the one hand, the capacity of a corpus to comprise the different textual types in a given variety or language and, secondly, its ability to account for variation within it. For the design of the *BLaRC*, which was created primarily to identify and analyse its legal terminology implementing different automatic methods, a decision was made to focus solely on law reports, given their relevance within the British legal system in comparison with other legal text types, as stated above. Furthermore, law reports touch upon all areas of law so the corpus was structured according to the field the texts pertained to so as to be able to account for terminological variation across legal areas.

Nevertheless, the question whether a specialised corpus is big enough to be representative of a given variety of the language, even if it is balanced and well sampled, still remains open to debate. There seems to be no clear agreement concerning the recommended size for a specialised corpus basically due to the fact that most approaches to this question are made on a theoretical basis. Whereas Pearson (1998) proposes a million words as a reasonable number (she poses that the limit should rather be established by the number of texts available and convertible into digital format), Sinclair (1991) believes that corpora must be as large as possible, establishing 10 to 20 million words as the recommendable target for a specialised one.

On the other hand, Kennedy (1998) does not consider that a big corpus necessarily represents the language better than a small one. In addition to this, Flowerdew underlines that the size of a specialised corpus necessarily depends on the aim the corpus has been designed for, given that "specialised corpora are constructed with an *a priori* purpose in mind" (Flowerdew, 2004: 25). Nevertheless, only a few authors draw their

conclusions in this respect from actual data. Heaps (1978), Sánchez & Cantos Gómez (1997) or Corpas Pastor & Seghiri Domínguez (2010, citing Young-Mi 1995) propose measures to try and determine the most suitable size for a corpus.

Regarding the size of the *BLaRC*, an *a priori* decision had to be made for its compilation, since finding out about such data as type/token or term/type ratios to establish a word target based on actual data would require the existence of the corpus itself prior to its processing. Consequently, and following Biber's criteria on sampling and Sinclair's recommendations on specific corpus size, the initial target was set at 8.5 million words. As described in section 2.3, there were other external criteria which conditioned the structure and content of the corpus itself.

Following Sánchez & Cantos Gómez' (1997) study, which aims at formulating a method to try to determine the optimum size for a corpus to be representative of given language variety based on how the type/token ratio[4] progresses as the corpus grows bigger, type/term increase was measured in the *BLaRC*. Finding out the proportion of new terms appearing in a corpus as its size augments might be an objective way of determining whether the size of that corpus would suffice to study its terminology, as is the case with the *BLaRC*.

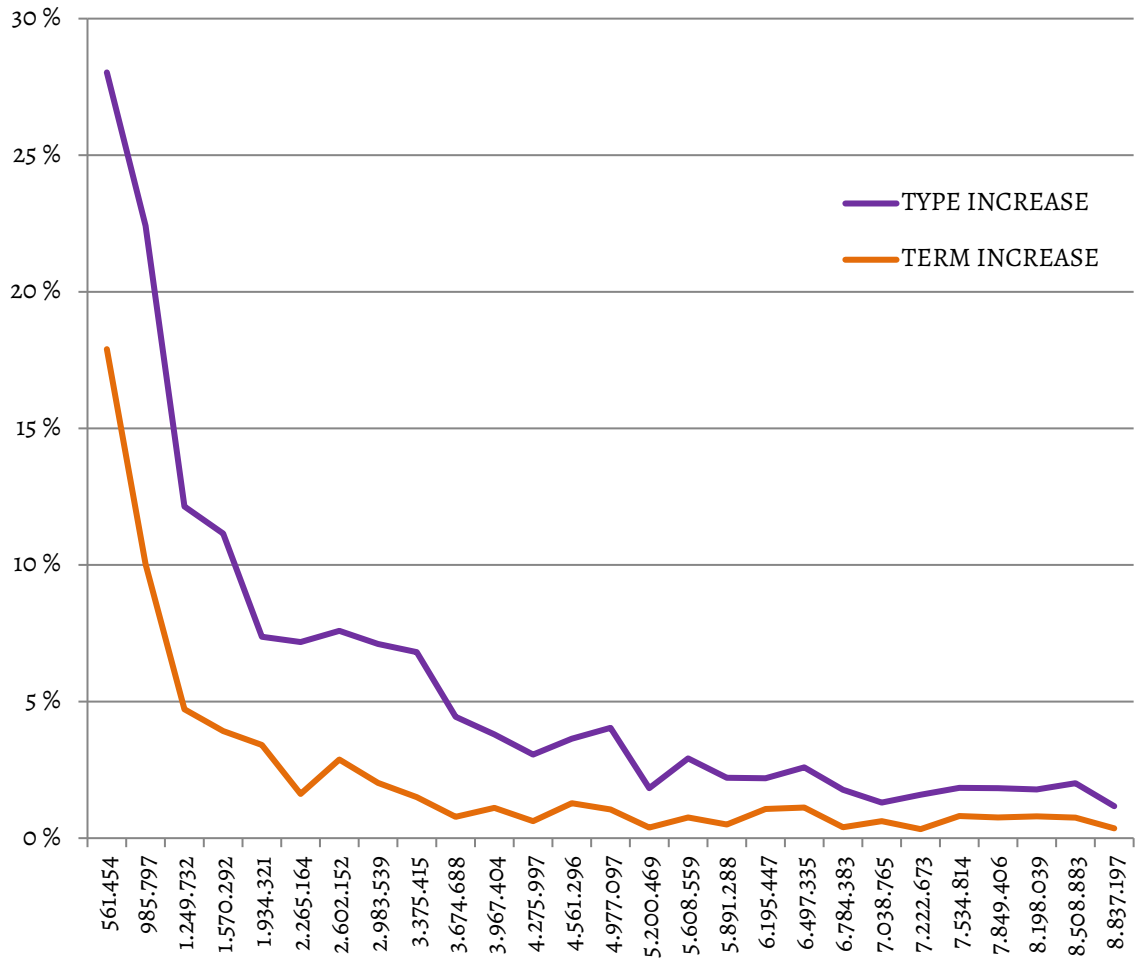
The terms in the *BLaRC* were first extracted automatically using Drouin's software *TermoStat* (2003) and then validated by comparison with a specialised legal English glossary of 10,088 terms.[5] Both the glossary and the lists generated by *TermoStat* (after progressively bringing together the 27 sub-corpora the main corpus was divided into) were compared using an excel spreadsheet so as to find out how many new terms appeared as new sub-corpora were added to the main corpus. The graph in Figure 1 illustrates the type/term ratio in the *BLaRC*, that is, how the percentage of terms and types, on the y-axis, relates to the total number of tokens in it. As can be observed, the former is inversely proportional to the latter, on the x-axis.

Figure 1 clearly illustrates how types and terms behave similarly, reducing their number as the corpus augments its size. Concerning the proportion of new terms appearing as the corpus grows bigger, they experiment a dramatic drop of 12.3 points from 17 % to 4.7 % as the corpus doubles its size from 500.00 words to 1.2 million approximately. Once the corpus reaches 1.2 million tokens, the decrease of new terms is less sharp falling from 10.03 % to 4.72 %. From that point on, although slightly recovering, this percentage drops to 1.62 % for sub-corpora 1 to 7 (2.26m tokens). It remains constant at 1.02 % on average until the corpus grows to 6.78 million words, decreasing to 0.4 % and not experimenting any significant changes from that point on.

[4] *Types* could be defined as the different words found in a corpus and the tokens associated to them through the type/token ratio coefficient are the repetitions of the same word within that corpus.

[5] Merged from four online legal glossaries available at www.legislation.gov.hk/eng/glossary/homeglos.htm, www.judiciary.gov.uk/glossary, sixthformlaw.info/03_dictionary/index.htm, and www.nolo.com/dictionary.

Figure 1: Type/term increase in the *BLaRC*.



Note: The x-axis represents the number of tokens.

Judging from the above, it appears that the initial target established for a corpus like the *BLaRC* may suffice to attain the objectives set for its compilation, that is, to analyse its terminology applying different automatic text analysis tools. As a matter of fact, 2.6 million words would have been enough due to the low increase in the percentage of new types and terms appearing as the corpus grew bigger. This is the reason why a pilot corpus of that size (*The United Kingdom Supreme Court Corpus*) was extracted from the *BLaRC* in order to facilitate the implementation of the methods described in section 3 and the analysis of the data.

2.3. Distributional criteria and word targets per category

The number of texts comprised by the *BLaRC* is not evenly distributed amongst its categories (which follow the geographic and hierarchical distribution of the courts and tribunals in the UK). Great variation was found depending on the text source (court or

tribunal[6]). The reasons for the irregular distribution of the texts available are varied, in some cases, especially regarding tribunals, they may have started working recently or disappeared due to the *Tribunals, Courts and Enforcement Act, 2007*. In some others, the high figures coincide with a densely populated area (one of the criteria supporting text distribution within the corpus) or with a court whose decisions, due to its high status in the hierarchy (i.e. any of the chambers of the High Court of Justice of England and Wales), set binding precedent and may thus be more relevant for legal practitioners when it comes to arguing a case.

Nevertheless, the targets established for the sections and subsections of the corpus were kept proportional to the total number of texts available within the covered time span. Subsequently, the sub-targets were set according to this criterion: if the number of texts in a section was higher, they were assigned a larger word target, thus being more representative of the language variety as that is the proportion they keep in real life, or at least this was assumed to be so.

These decisions were made following Biber's (1993; 1998) recommendations so as to ensure the ability of a corpus to represent a variety of the language properly. When designing the corpus itself, researchers should bear in mind variability, which "can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness" (Biber, 1993: 247). The geographical and institutional criteria that influenced the structure of the corpus above might fall within the "situational" perspective, according to Biber, whereas thematic and terminological criteria could be classified as linguistic.

All the same, a corpus should not be intended to systematise reality in a mathematical way, in this case, we simply intended to be as coherent as possible in every step we took towards corpus design. As Sinclair puts it when dealing with the issue of sampling a corpus and the structural criteria to employ when designing it: "real life is rarely as tidy as this model suggests" (Sinclair, 2005: 3). Moreover,

> "We remain (...) aware that the corpus may not capture all the patterns of the language, not represent them in precisely the correct proportions. In fact, there are no such things as "correct proportions" of components of an unlimited population" (Sinclair, 2005: 4).

Having said so, the total number of texts available between 2008 and 2010 was 16,612. Therefore, the word targets were established with respect to it, as already stated. Table 1 shows how this distribution was organised for the section devoted to those texts coming from English and Welsh institutions, by showing the total number of texts available per sub-category, their percentage with respect to the total amount of texts and the corresponding word target achieved following this proportion.

[6] Note that "the essential difference between a tribunal and a court is that a tribunal does not administer any part of the 'judicial power of the state'. It has a specific jurisdiction as allocated by Parliament and does not enjoy a broad jurisdiction defined in general terms" (Geary, 2012: 51).

Table 1: England and Wales courts and tribunals.

Court / Tribunal	available texts	% of total	final word target
Court of Appeal (Civil Division)	2,640	15.89 %	956,398
Court of Appeal (Criminal Division)	1,136	6.84 %	414,683
High Court (Administrative Court)	2,039	12.27 %	731,693
High Court (Admiralty Division)	17	0.11 %	8,842
High Court (Chancery Division)	1,009	6.07 %	366,298
High Court (Commercial Court)	379	2.28 %	142,701
High Court (Court of Protection)	26	0.16 %	34,007
High Court (Senior Costs Off.)	70	0.43 %	29,302
High Court (Family Division)	199	1.20 %	84,557
High Court (Mercantile Court)	8	0.05 %	6,152
High Court (Patents Court)	105	0.64 %	40,420
High Court (Queen's Bench Division)	709	4.27 %	255,301
High Court (Technology and Construction Court)	284	1.71 %	101,066
Patents County Court	12	0.08 %	15,242
Magistrates' Court (Family)	98	0.59 %	33,680
County Court (Family)	56	0.34 %	20,702
Care Standards Tribunal	70	0.43 %	27,762
Lands Tribunal	115	0.70 %	44,004
Total	**8,972**	**54.06 %**	**3,322,810**

Note: The final word target was obtained by calculating the number of words which the percentage displayed in the third column represented with respect to the initial word target, 8.5 million words. In order not to include truncated versions of some of the decisions in each section, the final word target sometimes exceeded the expected size slightly, respecting the actual length of the decisions comprised in the corpus.

3. Applications of CL techniques to the study of *legalese*: Automatic Term Recognition methods

Once the corpus has been properly compiled and structured, the applications to the study of the language samples comprised in it are manifold. Amongst other, we find Automatic Term Recognition (ATR henceforth). Yet, as stated in research question number 2: How can ATR methods contribute to the study of legal texts? Can we trust these methods as dependable tools to rely on?

To begin with, ATR methods can become extremely useful tools for the researcher interested in handling large amounts of information that could not be processed manually. In fact, getting to know the most significant terms in a corpus of specialised

texts can definitely contribute to a better understanding of the texts themselves, since terms could be defined as "linguistic representations of domain-specific key concepts in a subject field that crystallise our expert knowledge in that subject" (Kit & Liu, 2008: 204) and also lead to the identification of relevant topics that would otherwise remain unnoticed. In sum, specialised terms could be regarded as conceptual vehicles which can be employed to transmit specialised knowledge amongst scientists, researchers, or professionals in all specialised areas, hence their relevance and the need to identify them within a text collection. Actually, mining the specialised terms from a text collection might be the point of departure for further enquiry into the texts in a corpus by focusing, for instance, on collocate patterns (either as pairs of collocates of collocate networks), as shown in the last sections.

In order for ATR methods to be trusted as useful tools for term mining, and given the peculiar statistic behaviour of legal terminology, it becomes necessary to test them in order to select the most efficient ones in legal term extraction. It is commonly acknowledged that legal English is deeply intertwined with general language (Alcaraz, 1994; Borja, 2000; Mellinkoff, 1963; Tiersma, 1999), displaying specific features such as the abundance of sub-technical terminology, in other words, of "common words with uncommon meanings", (Mellinkoff, 1963) whose frequency and distribution might often be similar in the general and specific fields. ATR methods resorting to corpus comparison employ such parameters as frequency and distribution to perform their function. If a given term behaves similarly (in statistical terms) in both contexts, an ATR method implementing corpus comparison may be likely to fail or be less efficient and produce output lists of candidate terms that might contain a high percentage of noise (of false terms).

Consequently, ATR methods must be tested so as to identify the most effective ones in legal term recognition. In the past, the literature on ATR methods and software tools has been profusely reviewed (Maynard & Ananiadou, 2000; Cabré Castellví, Estopà Bagot & Vivaldi Palatresi, 2001; Drouin, 2003; Lemay, L'Homme & Drouin, 2005; Pazienza, Pennacchiotti & Zanzotto, 2005; Chung, 2003; Kit & Liu, 2008 or Vivaldi et al., 2012, to name but a few) often classifying these methods according to the type of information used to extract candidate terms (CT) automatically. One of the research foci of these works is the level of efficacy such methods can reach, concentrating on the amount of true terms (those terms confirmed as such after validation) they are capable of identifying automatically. In general, the most widespread procedure to determine the efficacy of ATR methods consists in comparing the list of CTs identified by each of them against a gold standard, that is, a glossary of specialised terms which ATR method designers employ as reference.

In Marín (2014; 2015) we find the evaluation of ten different ATR methods leading to the identification of the most efficient ones in the legal field. Table 2 displays the rate of efficiency reached by those ATR methods devoted solely to single-word term recognition. The figures show that it is Drouin's (2003) method which manages to success-

fully extract a greater rate of legal terms both on average (73 % of the terms identified were confirmed as true terms) and also for the top 200 candidate terms in the output lists (88 % of these were confirmed as legal terms).

Table 2: Average precision reached by SWT recognition methods (Marín, 2015: 11).

ATR Method	Avg. Precision 2,000 CTs	Precision Top 200 CTs
TermoStat (Drouin, 2003)	73.0 %	88.0 %
Kit and Liu (2008)	64.0 %	84.0 %
Keywords (Scott, 2008)	62.0 %	85.0 %
TF/IDF (Sparck Jones, 1972)	57.4 %	74.5 %
Chung (2003)	42.5 %	48.5 %

Note: ATR = Automatic Term Recognition; CT = Candidate Term.

The assessment process carried out by Marín (2014; 2015) consisted in the automatic validation of the candidate term lists produced by each method against a legal English glossary used as gold standard (see footnote 5 on the description of the glossary). The output lists were compared with the gold standard using an excel spreadsheet with the aim of determining the overlap percentage existing between both lists. Whenever a candidate term was found in the glossary, it was confirmed to be a true term. Therefore, the percentages found in the table above could be interpreted as the average level of precision achieved by each of the evaluated methods.

As regards Drouin's *Termostat* (2003), it is based on previous work on lexicon specificity such as Muller's, Lafon's, or Lebart & Salem's (in Drouin, 2003). Drouin claims that the frequency of technical terms in a specialised context differs, in one way or other, from the same value in a general environment and that "focusing on the context surrounding the lexical items that adopt a highly specific behaviour [...] can help us identify terms" (Drouin, 2003: 100). This author uses a corpus comparison approach which provides information on a candidate term's standard normal distribution giving

> "access to two criteria to quantify the specificity of the items in the set [...] because the probability values declined rapidly, we decided to use the test-value since it provides much more granularity in the results" (Drouin, 2003: 101).

Drouin applies human and automatic validation methods to evaluate the levels of precision and recall of his method. The author also resorts to three specialists who identify the true terms (TT) from the list generated by *TermoStat* noticing that subjectivity played a relevant role in this evaluation phase and that it might also be interesting to study human influence on validation processes. Regarding automatic validation, he compares the lists of CTs with a telecommunications terminology database. *TermoStat* reaches 86 % precision in the extraction of SWTs.

The ATR method designed by Drouin (2003) offers a user-friendly online interface,[7] which allows the researcher to upload their corpus (it accepts French, English, Spanish, Italian and Portuguese texts) and process it easily, obtaining the ranked list of candidate terms and other useful information for the analysis of the terminology comprised in it. Once the corpus is processed (it allows for the upload of files up to 30 megabytes), *TermoStat* produces a list of lemmatised[8] terms which are ranked according to their level of specialisation. Drouin's method resorts to corpus comparison for term extraction, using a reference corpus of newspaper articles as the general language corpus.

Figure 2: Output list of candidate terms extracted by *TermoStat*.

Candidate (grouping variant)	Frequency	Score (Specificity)	Variants	Pattern
section	9694	126.29	*section* *sections*	Common_Noun
v	6828	112.55	*v*	Common_Noun
case	11465	111.79	*case* *cases*	Common_Noun
para	5973	108.63	*para* *paras*	Common_Noun
article	5686	97.39	*article* *articles*	Common_Noun
court	6387	88.65	*court* *courts*	Common_Noun
appeal	3993	80.3	*appeal* *appeals*	Common_Noun
appellant	3102	78.47	*appellant* *appellants*	Common_Noun
not	22062	75.07	*not*	Adverb
law	5484	73.55	*law* *laws*	Common_Noun
judgment	2862	71.67	*judgment* *judgments*	Common_Noun
claim	3293	69.8	*claim* *claims*	Common_Noun
right	5795	67.98	*right* *rights*	Common_Noun
apply	3542	65.5	*apply* *applying*	Verb

As shown in Figure 2 the output list includes not only is the term's specificity value (spécificité) but also its frequency as lemma (fréquence), its variants (variants orthographiques), and its part-of-speech tag (matrice). The lexical categories identified by *TermoStat* are: nouns, adjectives, adverbs and verbs. It also detects multi-word terms having nouns and adjectives as phrase heads.

Table 3 displays the top 25 candidate terms (prior to the validation of the method) as ranked by *TermoStat* according to its level of specialisation, or specificity level, that is, after implementing the algorithm designed by the author. As it can be observed in the table below, not all the terms identified by the system could be regarded as legal terms proper. As already stated, this table includes all the candidate terms Drouin's method managed to extract before the whole list was validated against our legal glossary. We decided to offer this data for the reader to acknowledge the possibilities at hand using

[7] Online at http://termostat.ling.umontreal.ca.

[8] The term *lemma* refers to the root word without any inflectional suffixes (for instance, the infinitive of a verbal form). Lemma frequency includes all the occurrences of any of the possible realisations of the root word. Those methods which resort to lemmatisation tend to be more efficient than those which do not.

this term extraction method, which managed to identify 88 % legal terms out of the top 200 candidate terms extracted automatically from the *BLaRC*.

Table 3: Top 25 terms as identified by Drouin's *TermoStat*.

Rank	Term	Specificity level	Rank	Term	Specificity level
1	section	126.29	14	order	64.39
2	v (versus)	112.55	15	decision	63.53
3	case	111.79	16	person	62.83
4	para (paragraph)	108.63	17	proceeding	61.70
5	article	97.39	18	relevant	59.02
6	court	88.65	19	purpose	58.45
7	appeal	80.30	20	defendant	57.72
8	appellant	78.47	21	provision	57.55
9	law	73.55	22	principle	55.77
10	judgment	71.67	23	application	55.50
11	claim	69.80	24	jurisdiction	55.50
12	right	67.98	25	paragraph	54.69
13	apply	65.50			

4. Term collocates and lexical networks: Williams (2001) and Brezina, McEnery & Wattam (2015)

Closely linked to the automatic identification of specific terms is the relevance, not only of the terms themselves, but also of other words which tend to co-occur with them, that is, their collocates. Yet, going back to the research questions posed in the introduction, how can such patterns contribute to the study of legal text? Are there any automatic tools which facilitate such task?

Collocational patterns reveal the context in which a word occurs and provide plenty of information about the meanings and connotations associated with a word in context. When it comes to sub-technical or polysemous terms, their collocates can also help us distinguish between their specialised and general meaning but, most importantly, can point at other questions that may remain unnoticed on a superficial reading of legal texts. Nevertheless, for the identification of collocational patterns in a text collection, especially if it is a large corpus, it is necessary to employ automatic tools that facilitate the task. Let us first define and consider some theoretical questions re-

lated to the concept of collocation and then move onto the actual usage of collocation extraction software and its applications to the study of legalese.

Broadly speaking, in Firth's words, a collocate is "the company a word keeps" (1957: 6). The concept *collocation* has been revisited since then (Cruse, 1986; Gries, 2013; Sinclair, 1991; Stubbs, 2001) and more specific and accurate definitions have been provided, John Sinclair's being a classic reference in the field. Sinclair (1991; 2005) deems the statistical data associated with two co-occurring words as fundamental for their identification, as collocates can be mined automatically by applying measures of association like mutual information (Church & Hanks, 1990) or log-likelihood (Dunning, 1993), amongst others. Williams elaborates on this idea by delimiting the concept of collocation as

> "the habitual and statistically significant relationship between word forms within a predefined window and for a defined discourse community, expressed through an electronic corpus of texts" (2001: 5).

On a semantic level, based on the work by Stubbs (2001) on semantic preference and discourse prosody, Baker (2016: 2) insists on the mutual influence that collocates have on each other as regards their meaning, affirming that "collocates help to imbue words with meaning as words can begin to take on aspects of the meaning of the words that they collocate with".

However, as Baker (2016) acknowledges, the study of collocates has been limited to the analysis of word pairs until recently, often due to the limitations of tools like *AntConc* (Anthony, 2014) or *Wordsmith* (Scott, 2008), only capable of extracting pairs of collocates, disregarding the potentiality of collocational or lexical networks (Williams, 2001) in the study of the interaction amongst terms and their vicinity in a corpus.

Geoffrey Williams (2001) is one of the first authors to explore word associations beyond word pairs in specialised contexts based on the work by Phillips (cited in Williams, 2001). Williams proposes the lexical network model, which puts forward a quantitative approach to the study of word usage through the analysis of their collocates and co-collocates. The context is thus extended since lexical networks, which revolve around a central word or node, spread out progressively by also including the node's co-collocates and, in turn, the collocates of those co-collocates.

Williams' (1998) idea that collocational or lexical networks may enhance quantitatively and, above all, qualitatively our understanding of specialised vocabulary meant a step forward in the study of term usage and meaning and authors like Baker (2005; 2016), McEnery (2006) or Marín (2016) acknowledge this fact. However, in spite of the above, the process undergone in the production of lexical networks could be time consuming, as Baker (2016) and Marín (2016) affirm, requiring the manual arrangement of the networks (often populated by thousands of elements), since automatic corpus tools only allow for the study of one collocational level.

There is a plethora of tools capable of processing electronic text designed with different purposes (Sternfeld, 2012) although not many of them can obtain the lexical

networks of a term automatically. This is the case of *Voyant Tools* (Sinclair et al., 2012) and *Lancsbox* (Brezina, McEnery & Wattam, 2015). Both offer plenty of possibilities to exploit corpora. The former is extremely powerful in loading large amounts of text online and offers very visual applications like *Cirrus*, *ScatterPlots* or *TermsRadio*, amongst other. Nevertheless, as regards collocate networks, the proposal by Brezina, McEnery & Wattam's (2015) proposal appears to be grounded and motivated by more solid linguistic criteria, allowing for a deeper analysis of the collocate networks of terms. It goes further than *Voyant Tools* into the contexts of usage not only of the central node of the networks but also of its collocates and co-collocates. Furthermore, *Lancsbox* implements the possibility of modifying the measures applied to obtain a word's collocates and thus test the efficacy of the tool in producing relevant collocate inventories, depending on the users' preferences.

One of the advantages of using *Lancsbox*[9] is that it not only manages to obtain a word's network very quickly, but also visually represents the network through a graph that displays the node's collocates, connecting them with vectors whose size varies according to the strength of the collocational bond calculated by the tool (the shorter the vector, the stronger the link between words) and indicating collocate directionality. *Lancsbox* also presents the possibility of adjusting association measures by testing which one produces the most interesting results. Amongst other, measures such as *MI3*, *delta-p* or *log-likelihood* can be implemented in the production of a word's lexical network, represented by a graph, as shown below.

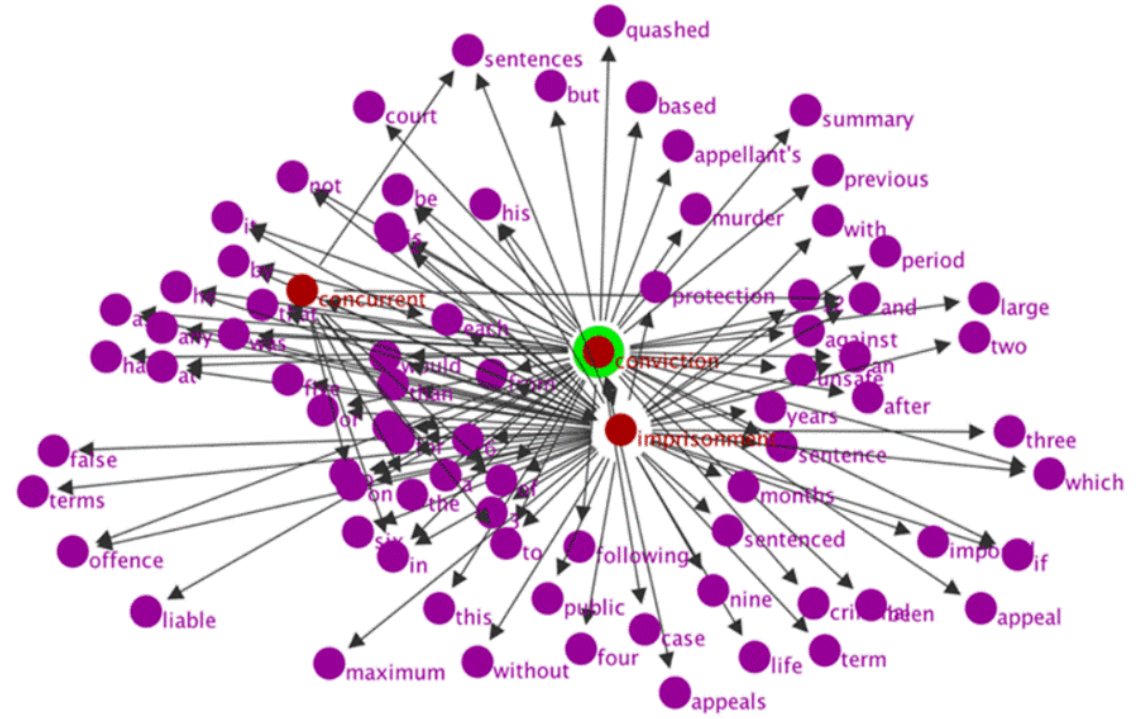
Once they are obtained, the graphs contain detachable tabs, which permit the user to generate embedded collocate networks, always displaying the relationship amongst all their constituents and the main node, as illustrated by Figure 3. If we click on any of the collocates (in purple), a new collocational level will be shown, which includes the collocate's collocates, that is, those words which tend to co-occur with each of the node's collocates. This can be done up to seven times, thus allowing for a subsequent development of the networks to the seventh collocational level.

As shown in Figure 3, which displays the collocational network of the term *conviction* (circled in green), it presents first level collocates such as *imprisonment*, *summary*, *appeal* or *sentence*. If we had not resorted to *Lancsbox*, the collocational network would have stopped at this point, however, this tool enlarges the context by displaying the collocates of *imprisonment* (in red), namely, *concurrent*, *conviction*, *sentence* or *protection* and of those words which also collocate with it, such as *concurrent* (the third sub-node, which constitutes the third collocational level in the network below). Whenever any of these share any collocates, they are linked with an arrow which indicates collocate directionality. Owing to the fact that the corpora employed in this study are considerably large (13.7 and 8.5 million words respectively), the networks might be excessively populated, as displayed in Figure 4. This is why the frequency thresholds must be adjusted to pre-

[9] Available at http://corpora.lancs.ac.uk/lancsbox/index.php.

vent this from happening. In any case, the tables appearing to the left of the graphs (as shown in Figure 4), once they are generated, allow the user to navigate through the whole collocate inventory easily.

Figure 3: Specialised collocational network of the word *conviction* (in *BLaRC*).



One of the advantages of *Lancsbox* is the possibility of adjusting the settings to limit the number of collocates in the networks or to change the association measures employed to mine them, as already stated. This is why Brezina, McEnery & Wattam (2015) perform a case study analysis where different measures are used in the replication of McEnery's (2006) examination of swearing language (the words *swearing* and *drunkenness* exemplify the study). In spite of all the multiple applications and advantages of *Wordsmith* (Scott, 2008), the software McEnery uses to extract the collocates in his study, it does not offer the possibility to implement MI3 (the cubed version of Church & Hanks' (1990) mutual information measure). In a nutshell, what mutual information does is basically to compare

> "the probability of observing *x* and *y* together (the joint probability) with the probability of observing x and y independently (chance). If there is a genuine association between x and y, [...] then the joint probability will be much larger than chance" (1990: 77).

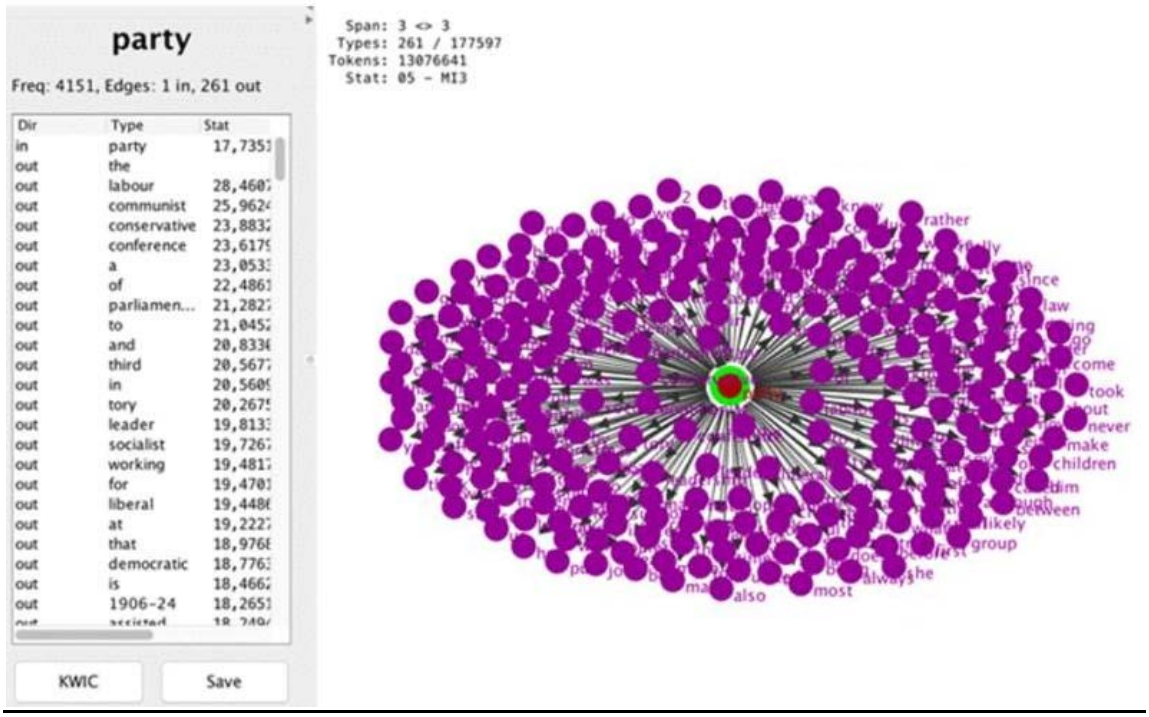
Therefore, if a collocate pattern was assigned a high MI score owing to its joint statistical behaviour, it would be identified as relevant within a given text collection.

As already stated, McEnery opts for mutual information (MI), highly precise, although it often shows a certain "propensity to highlight unusual combinations [...] that co-occur only once or twice in the corpus" (Brezina, McEnery & Wattam, 2015: 159). A

collocate frequency threshold would thus become necessary for the networks not to become unmanageable and excessively populated if MI was to be applied. On the contrary, MI3 tends to push more frequent combinations to the top of the rank, leaving the most unusual patterns aside or either relegating them to the bottom of the collocate inventories, in other words, "the measure gives more weight to observed frequencies and thus gives high scores to collocations which occur relatively frequently in the corpus" (Brezina, McEnery & Wattam, 2015: 160).

The data associated with each of the constituents of the network can also be read in detail and saved in .csv format. The extension .csv stands for "comma separated values", which can be easily imported into an excel spreadsheet. As seen in Figure 4, a table displays the collocates of the selected item (highlighted in green in the graph) and also the value assigned to each pattern by the algorithm implemented through MI3 together with the raw and relative frequency of each pattern on the list.

Figure 4: *Lancsbox* table and graph as shown by the interface control panel.



Having said this and leaving aside the fact that *Lancsbox* is capable of producing the lexical network of a term on the fly, which, on its own, is a major improvement, Brezina, McEnery & Wattam emphasise that the main potential of this software is its capability to unveil the semantic interaction amongst the words in a corpus by extending a word's context beyond the word itself and avoiding the painstaking and time-consuming process of doing it manually, as Baker (2016) and Marín (2016) also acknowledge.

5. Subtechnical legal terms and collocational networks: A case study

Following from the above, the applications of *Lancsbox* to the analysis of corpora and their lexicon are manifold. As Marín (2016) demonstrates in the proposal of an algorithm to study the level of specialisation of subtechnical vocabulary, the relevance and significance of this particular type of legal terminology in a corpus of judicial decisions was considerable. The comparison between the list of specialised legal terms extracted from the *British Law Report Corpus* and the list of the 3,000 most frequent words of English found in the *British National Corpus* (2007) yielded 45.41 % overlap, thus showing "that approximately half of the legal terminology identified in the *BLaRC* is shared with the general field, since almost 50 % of it matched the general vocabulary lists" (Marín, 2016: 81).

As shown in section 3, this is a common feature of the legal English lexicon, however, very little has been said about the meaning of these words in context. Words such as *trial*, *relief*, *battery* or *charge* (which are statistically profiled in Marín's analysis) present a specialised meaning in the legal context which very rarely occurs in the general one. Sections 5.1 and 5.2 present a case study illustrating the applications of Lancsbox to the study of subtechnical legal terms.

5.1. Methodology

Two corpora were employed in this analysis, one of them the *BLaRC* (8.5 million words), the other one *LACELL*, a 13.7 million word general English corpus containing texts from various British sources such as newspapers articles, book chapters (academic, fiction, etc.), magazine articles, brochures, letters and the like. Both corpora were processed using *Lancsbox* (Brezina, McEnery & Wattam, 2015). The thresholds established to limit the amount of collocates generated by the system were, firstly, >10 frequency, according to which, the pairs of collocates and co-collocates should co-occur at least 10 times in the corpus to be mined by the system. Secondly, the collocate window cut-off point was 3, that is, the collocates included in the network should fall within the three immediate words to the left and right of the node (the search word) or any constituent of the network. Following Brezina, McEnery & Wattam (2015) and Baker (2016), the association measure implemented for the calculation of the term's collocate network was MI3, whose capacity to leave irrelevant patterns aside by pushing them to the bottom of the collocate ranks has already been discussed.

The word selected for this case study is *party*, a sub-technical word whose presence in both corpora is remarkable, hence its sub-technical character, displaying 4,808 raw frequency in the general corpus (3.5 relative frequency) and 40 % distribution (it ap-

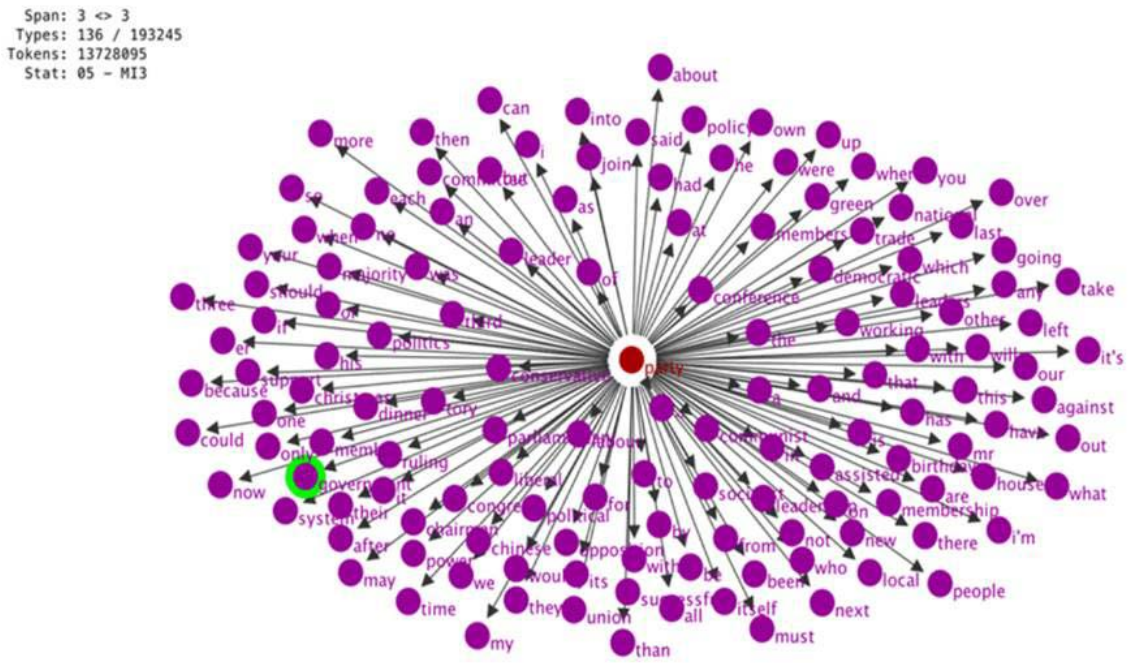
pears in 1,712 out of 4,281 texts). In contrast, its frequency in the specialised corpus is 4 points higher than the same value in the general corpus (if we compare their relative frequencies), as it occurs on 10,351 occasions (7.5 relative frequency). In addition, it presents higher distribution values, covering 73 % of the texts in it.

Nonetheless, the major difference found between the use of *party* in the general context and the specific field, as might be expected, is related to its meaning in both areas. It is at this point that the software package *Lancsbox* can provide evidence of the context which surrounds the term, establishing which of its meanings in each corpus is the most representative one. The collocates associated to each of the senses of the word *party* (the main node of the lexical network obtained with *Lancsbox*), illustrate how the meaning of *party* can be understood as a "political group" or "celebration" or acquire its legal sense in the specialised corpus, meaning "person/s taking part in a legal proceeding".

5.2. Results and discussion

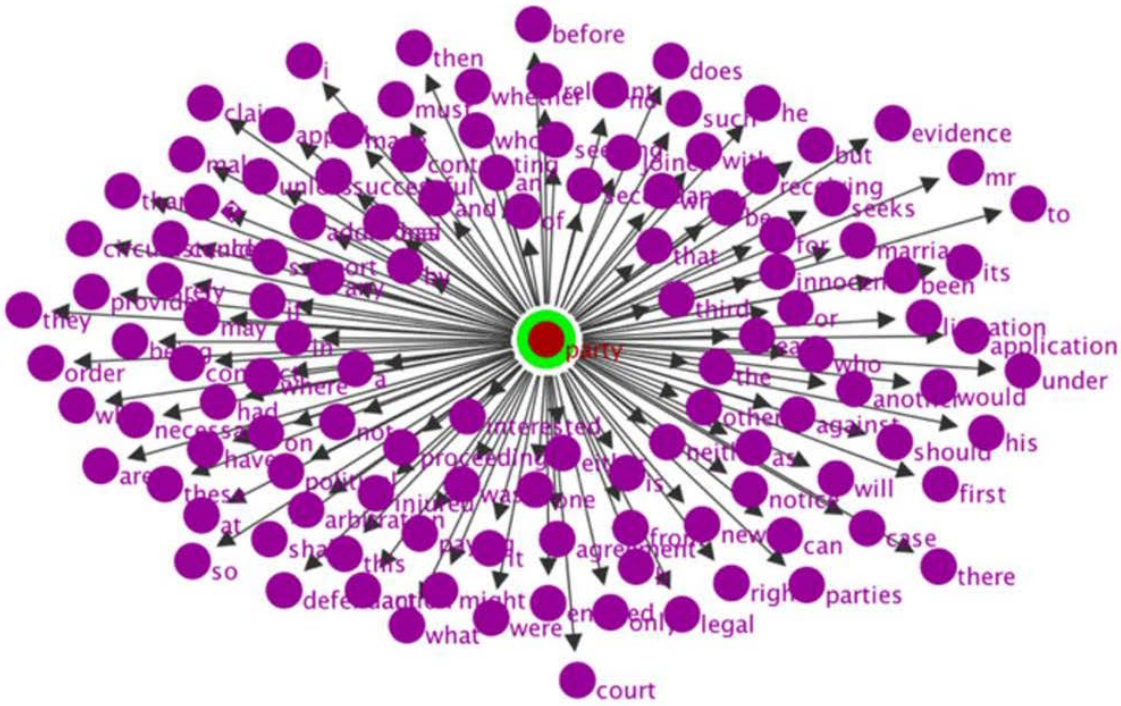
Figures 5 and 6 display the first level collocate networks of the term *party* in both the specialised and the general fields. In a first approach, and judging by the stronger lexical collocates of *party* in the general corpus (this is indicated by the shorter vector that joins them and by the coefficient displayed in the table attached to the graph, not in the figure), the primary meaning of the term is clearly "political group/association", in fact, the words *labour, communist, conservative, parliamentary, tory, leader or socialist* appear amongst the top 25 collocates identified by *Lancsbox.*

Figure 5: 1st level collocate network of *party* in the general corpus.



Evidence of its secondary general meaning, "celebration", much less frequent in *LACELL* than the former, was also found in the general corpus, as it was expected. Words like *dinner, cocktail, birthday, Christmas* or *tea* can be found within the top 125 collocates of *party*. On the other hand, its collocates in the legal corpus clearly signal its legal sense since we find words like *third, proceedings, innocent, agreement, contracting, marriage, aggrieved* or *arbitration* ranking amongst the top 50 collocates of the term.

Figure 6: 1st level collocate network of *party* in the legal corpus.



A detailed observation of the elements found in these networks can also help us identify certain topics related to the node (the main search word), which could be explored further by extending the network to a lower collocational level through the selection of any of the collocates in the set displayed above.

Government is one of them. If the collocates associated with it are examined on a second collocational level, still within the general English corpus *LACELL*, we observe a portrayal of this institution as reflected on texts coming from various sources such as the press, books, brochures, advertisements, written correspondence, etc. One of the possibilities of analysis could be grouping the network constituents according to semantic categories, as they refer to the different functions, organisation and features of this ruling body. The words *local, central, departments, federal* or *regional* belong in this area. On the other hand, the term *government* is associated with the ideology of the parties exercising that function, the collocates *tory, conservative* or *labour* are indicative of this fact. Another group of collocates which also contribute to the linguistic characterisation of this institution are those which refer to the power it exerts. Words like *reform, control* or *power* fall within this category. In addition, the word *government* occurs with

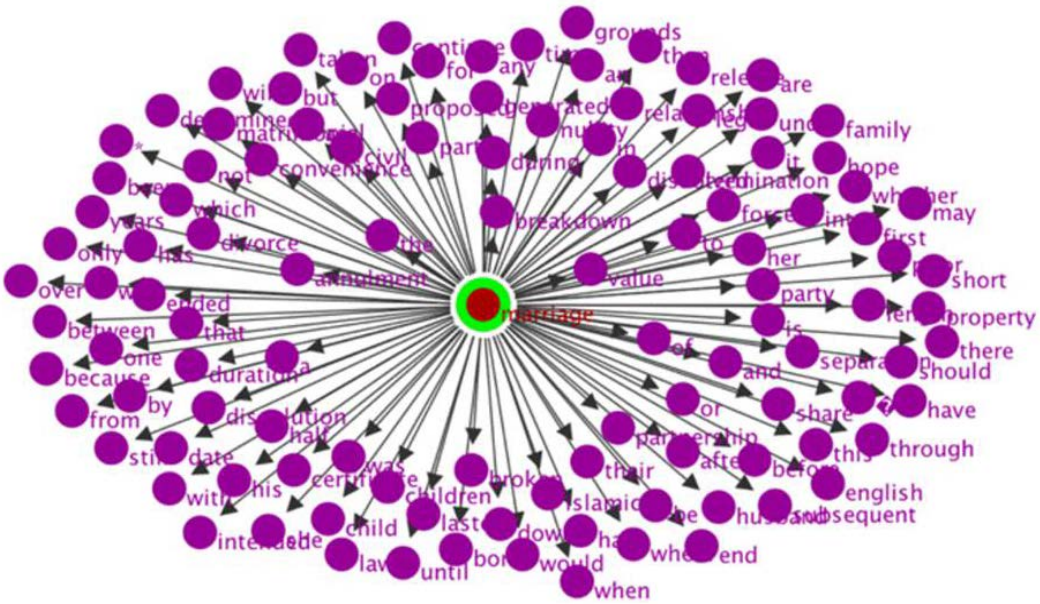
words related to its *public service* role being envisaged as an institution which acts as *guidance* in public matters, takes care of *people's welfare* and is at their *service* (the words in italics are also collocates of *government)*. Likewise, its collocates display a general concern about economic issues since the words *funding, spending, expenditure* or *taxes* appear in its lexical network. The concordance lines below attest how all these examples can be analysed and interpreted in context. *KWIC* (key words in context), a software utility included in *Lancsbox*, offers this possibility:

(…) he said that the GOVERNMENT would REFORM taxation (…)
(…) agreement must be struck between the CENTRAL and LOCAL GOVERNMENTS both on the central-bank system (…)
(…) The problem for this TORY GOVERNMENT is that their ideology is (…)
(…) interpreting the true spirit of GOVERNMENT GUIDANCE in plan making, (…) is but one consideration.
(…) seeing these as the main GOVERNMENT contributions to WELFARE, or the general good (…)
Rifkin told the Commons health SPENDING meant a third of the annual budget (…)

Concerning the legal context, the collocate network of *party* clearly reveals the legal sense of the term in the field, as expected. Other legal terms such as *proceedings, litigation, convention, liability* or *appeal* collocate with it as well as other words which, although not being used exclusively in the legal area, are associated with its legal meaning, namely, *contracting, marriage, innocent, financial* or *witness.* These collocates provide plenty of data on the nature of some of the cases which were brought before British courts between 2008 and 2010.

One of the words that caught our attention amongst the constituents of this network was *marriage*. The fact that an issue such as marriage might be so relevant as to rank in 30th position within the collocate inventory of *party* was interesting enough to delve into its lexical deployment in the legal corpus.

Figure 7: 2nd level specialised network of *marriage*.



The list of constituents of the lexical network of *marriage* is noticeably long, as shown by Figure 7, being also connected to a large number of collocates of the first level network node, *party*. According to their meaning, the most relevant lexical collocates of *marriage* point at two major elements of this relationship as reflected on the texts in the corpus. On the one hand, its legal character, on the other hand, the economic terms which the legal concept *marriage* revolves around. Amongst the former group we find *annulment, divorce, separation, civil,* or *nullity*. The latter category comprises words like *contract, value, banking, property, valuation* or *acquire*.

Within the group of collocates of the term *marriage,* the words *convenience* and *genuineness* caught our attention. According to the Immigration Act 1999 (sections 24 and 24 A), amended in this respect by the Immigration Act 2014 (section 55), a marriage of convenience is defined as a civil relationship where

> "one or both of the parties is not a British citizen […] there is no genuine relationship between the parties; either or both of the parties enter into the marriage […] for the purpose of circumventing immigration controls […]"

But how do these different aspects reflect on those judicial decisions where the collocate pattern *marriage of convenience* is employed? Firstly, we find several collocates which refer to the definition of the term itself as found in the law, namely, *sham, bogus, circumventing* or *genuine.* If we analyse the concordances of the collocate pattern *sham marriage* (which the law identifies with *marriage of convenience),* in an appeal to the Supreme Court by the Secretary of State for the Home Department of the UK, we find that

> "persons seeking leave to enter or remain in this country may marry here, not for the reasons which ordinarily and legitimately lead people to marry, but in order to strengthen their claims for leave to enter or remain. Such marriages have been variously described as 'bogus' or 'sham' and as 'marriages of convenience'."

The texts in the legal corpus also gathered sociological information in relation to the topic that may have remained unnoticed on a superficial analysis of a smaller text sample, unless we went deeper into the interconnections amongst the constituents of lexical networks at different levels. Words such as *prevalence, incidence, recurrence* or *usual* can be found amongst the collocates of the term *convenience,* which may lead us to explore the issue further by reading the concordances associated to these terms and exploring other references (newspapers, legal texts, journal articles) to support our findings in this respect.

Lastly, the second level collocate network of *convenience* also contains words and terms which point at the legal reaction to this phenomenon on the part of the legislative or executive bodies. As proved by data, marriages of convenience appear to be a significant judicial problem in the UK and words such as *prevent, supress, measures, fighting, battle* or *policing* may also be pointing at that fact. Let us observe in greater detail what the texts have to say about this issue:

> (…) it operates to PREVENT MARRIAGES of CONVENIENCE (…)

(…) section makes no reference to MARRIAGES of CONVENIENCE or SHAM MARRIAGES (…)
(…) MEASURES to be adopted on the COMBATING of MARRIAGES of CONVENIENCE (…)

In response to research question 3 on the usefulness of collocational patters in the study of legal text, this analysis has attempted to illustrate the multiple possibilities that the exploration of collocational networks offers to the researcher interested not only in the linguistic dimension of these texts but also in their legal or sociological one. The fact that these networks can be obtained easily by simply uploading a corpus using automatic processing tools like *Lancsbox*, simplifies the process enormously, since obtaining them semi-automatically requires lots of effort and time prior to the actual analysis of their content.

6. Conclusion

The present research has been conceived as an introduction into the design and compilation of legal corpora and their processing using automatic corpus analysis tools. Such introduction has been carried out through the description and processing of two corpora, a general one of 13.7 million words, *LACELL* – used as reference whenever a general English corpus was required for comparison – and *BLaRC*, a legal one of 8.5 million words, made up entirely of judicial decisions.

Concerning the first research question posed in the introduction, an effort has been made to highlight the relevance of sampling criteria in corpus compilation, focusing, on the one hand, on the communicative relevance of the texts in the corpus and on the other hand, on the structure of the corpus itself.

Firstly, law reports have been presented as a fundamental legal genre all legal practitioners must know and cite, hence their importance within this ESP variety. Secondly, as regards the structure of the corpus, such a controversial issue as establishing the ideal word target has been tackled, concluding that, after calculating the type/term ratio in our legal corpus, a 2.5 to 3 million word target could suffice to study its lexicon, since the proportion of terms per word type dropped drastically at that point. The general structure of our legal corpus has also been presented in section 2.3., where a proportion in the word targets for each corpus category and subcategory was kept according to the number of texts available for each of them.

The second research question in the introduction enquired about the usefulness of Automatic Term Recognition (ATR) methods in the analysis of legal text. As shown in section 3, ATR methods can be of great help to the researcher when handling large amounts of data which could not be processed otherwise. Terms encapsulate specialised meaning, however, not all automatic term recognition methods are equally efficient in legal term identification. One of the reasons that could account for this phenomenon is the close relationship between legal terms and everyday vocabulary, where

large percentages of the former can be found. This is why different ATR methods were tested in order to select the most efficient ones in the legal field. The result of the assessment of five different ATR methods has been presented in section 3. After the validation process, it was found that Patrick Drouin's *TermoStat* (2003) managed to identify correctly 73 % legal terms in the *BLaRC*, ranking first in legal term mining. *TermoStat* is therefore recommended as the best method to extract legal terminology, which often poses difficulties in the accomplishment of this automatic task, as already stated.

Finally, the third research question posed in the introduction has been answered in sections 4 and 5, where one of the latest trends in Corpus Linguistics has been presented, that is, the use of software tools for the examination of collocate networks. A case study has been carried out in section 5 using one of these tools: *Lancsbox* (Brezina, McEnery & Wattam, 2015). One of the advantages of exploring the collocate patterns in a corpus is that they are capable of bringing to the foreground relevant aspects of its content and form that may otherwise remain unnoticed. Thanks to *Lancsbox* the task of producing collocate networks can be accomplished on the fly, allowing for the deployment not only of a word's collocate network but also of the networks associated with its collocates and the collocates of those collocates up to a seventh hierarchical level. The possibilities of enlarging the context of usage of a given word and analysing it through such connections are manifold.

To conclude, section 5 has demonstrated how the meaning of the sub-technical term *party* radically changes from one context to the other and how those meanings are organised in a hierarchical way in both contexts. Such change has been observed through the analysis of the constituents of the collocate networks extracted from both corpora, which have shown how the prevailing sense of the term *party* in the general corpus was that of "political group/association", followed by "celebration", whereas it meant "person/persons taking part in a legal proceeding" in the legal corpus, as was expected. Moreover, the collocate networks were explored in greater detail revealing interesting data such as the incidence of a topic like *marriage* in a corpus of judicial decisions, which, in principle, might not appear to be so relevant for a text collection comprising decisions from the criminal and civil fields. In fact, this analysis has gone beyond the merely linguistic level entering the sociological/legal dimension and allowing for a deeper understanding of such phenomena. In its creators' own words:

> "collocation networks as an analytical tool have a large potential in a number of areas of linguistic and social research such as discourse studies, psycholinguistics, historical linguistics, second language acquisition, semantics and pragmatics, lexicogrammar, and lexicology" (Brezina, McEnery & Wattam, 2015: 165).

Nevertheless, further research still remains to be carried out, particularly in the legal field, to test and exploit the potential of collocate networks, which this research has intended to suggest.

References

Alcaraz Varó, Enrique (1994). *El inglés jurídico: textos y documentos*. Madrid: Ariel Derecho.

Anthony, Laurence (2014). *AntConc* (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from www.laurenceanthony.net.

Baker, Paul (2005). *Public Discourses of Gay Men*. London: Routledge.

Baker, Paul (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21 (2), 139–164. DOI: 10.1075/ijcl.21.2.01bak.

Bhatia, Vijay (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, Douglas (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8 (4), 243–57. DOI: 10.1007/978-0-585-35958-8_20.

Biber, Douglas, Conrad, Susan, & Reppen, Randy (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biel, Łucja & Engberg, Jan (2013). Research models and methods in legal translation. *Linguistica Antverpiensia*, 12, 1–11. Available at lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/316/225.

Borja Albí, Anabel (2000). *El texto jurídico en inglés y su traducción*. Barcelona: Ariel.

Breeze, Ruth (2015). Teaching the vocabulary of legal documents: a corpus-driven approach. *ESP Today*, 3 (1), 44–63. Available at www.esptodayjournal.org/esp_today_back_issues_vol4.html.

Brezina, Vaclav, McEnery, Tony & Wattam, Stephen (2015). A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20 (2), 139–173. DOI: 10.1075/ijcl.20.2.01bre.

British National Corpus (2007). BNC XML Edition version 3, distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at www.natcorp.ox.ac.uk.

Cabré Castellví, María Teresa, Estopà Bagot, Rosa & Vivaldi Palatresi, Jordi (2001). 'Automatic term detection: a review of current systems', in Bourigault, Jacquemin & L'Homme (Eds.), *Recent Advances in Computational Terminology* (53–87). Amsterdam: John Benjamins. DOI: 10.1075/nlp.2.04cab.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Boston: The Massachusetts Institute of Technology (MIT).

Chung, Teresa (2003). A corpus comparison approach for terminology extraction. *Terminology*, 9 (2): 221–246. DOI: 10.1075/term.9.2.05chu.

Church, Kenneth Ward & Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16 (1), 22–29. Available at dl.acm.org/citation.cfm?id=89095.

Corpas Pastor, Gloria & Seghiri Dominguez, Míriam (2010). *El concepto de representatividad en lingüística de corpus: aproximaciones teóricas y consecuencias para la traducción*. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Cruse, David Alan (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Danet, Brenda (1980). Language in the Legal Process. *Law and Society Review, 14* (3), 445–564. DOI: 10.2307/3053192.

Drouin, Patrick (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9 (1): 99–117. DOI: 10.1075/term.9.1.06dro.

Dudley-Evans, Tony & St John, Maggie Jo (1998). *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.

Dunning, Ted (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19 (1), 61–74. Available at dl.acm.org/citation.cfm?id=972454.

Firth, John Rupert (1957) *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Flowerdew, Lynne (2004). The argument for using English specialised corpora to understand academic and professional language. In Connor & Upton (Eds.), *Discourse In The Professions: Perspectives From Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/scl.16.02flo.

Flowerdew, Lynne (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics* 14 (3), 393–417. DOI: 10.1075/ijcl.14.3.05flo.

Geary, Adam & Morrison, Wayne (2012). *Common Law Reasoning and Institutions*. London: University of London.

Goźdź-Roszkowski, Stanisław & Pontrandolfo, Gianluca (2014). Legal phraseology today: corpus-based applications across legal languages and genre. *Fachsprache: International Journal of Specialized Communication*, 3–4, 130–138.

Gries, Stefan Thomas (2013). 50-something years of work on collocations: What is or should be next. *International Journal of Corpus Linguistics*, 18 (1), 137–166. DOI: 10.1075/ijcl.18.1.09gri.

Gries, Stefan Thomas & Wulff, Stephanie (Eds.) (2010). *Corpus-linguistics applications. Current studies, new directions*. Amsterdam/New York: Rodopi.

Heaps, Harold Stanley (1978). *Information Retrieval: Computational and Theoretical Aspects*. New York: Academic Press.

Kennedy, Graeme (1998). *An introduction to corpus linguistics*. New York: Longman.

Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel & Suchomel, Vít (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1, 7–36. DOI: 10.1007/s40607-014-0009-9.

Kit, Chunyu & Liu, Xiaoyue (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14 (2), 204–229. DOI: 10.1075/term.14.2.05kit.

Lemay, Chantal, L'Homme, Marie-Claude & Drouin, Patrick (2005). Two Methods for Extracting 'Specific' Single-word Terms from Specialised Corpora: Experimentation and Evaluation. *International Journal of Corpus Linguistics*, 10 (2), 227–255. DOI: 10.1075/ijcl.10.2.05lem.

Maley, Yon (1994). The Language of the Law. In J. Gibbons (Ed.), *Language and the Law*. London: Longman.

Marín, María José (2014). Evaluation of five single-word term recognition methods on a legal corpus. *Corpora*, 9 (1), 83–107. DOI: 10.3366/cor.2014.0052.

Marín, María José (2015). Measuring precision in legal term mining: a corpus-based validation of single and multi-word term recognition methods. *ESP World*, 46, 1–23. Available at www.esp-world.info/Articles_46/MARIN_MEASURING%20PRECISION%20IN%20LTM-AN.pdf.

Marín, María José (2016). Measuring the degree of specialisation of sub-technical legal terms through corpus comparison: a domain-independent method. *Terminology*, 22 (1), 80–102. DOI: 10.1075/term.22.1.04mar.

Marín, María José & Rea Rizzo, Camino (2012). Structure and design of the BLRC: a legal corpus of judicial decisions from the UK. *Journal of English Studies*, 10, 131–145. DOI: 10.18172/jes.184.

Maynard, Diana & Ananiadou, Sophia (2000). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing* 8 (1), 101–125. DOI: 10.5715/jnlp.8.101.

McEnery, Tony (2006). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Abington, UK: Routledge.

McEnery, Tony & Wilson, Andrew (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony, Xiao, Richard & Tono, Yukio (2006). *Corpus-based language studies: an advanced resource book*. Routledge Applied Linguistics: New York.

Mellinkoff, David (1963). *The Language of the Law*. Boston: Little, Brown & Co.

Nesi, Hillary & Gardner, Sheena (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Orts Llopis, María Ángeles (2006). *Aproximación al discurso jurídico en inglés: las pólizas de seguro marítimo de Lloyd's*. Madrid: Edisofer.

Orts Llopis, María Ángeles (2009). Legal genres in English and Spanish: some attempts of analysis. *Iberica*, 18, 109–130. Available at www.aelfe.org/documents/07_18_Orts.pdf.

Partington, Adam (1998). *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Pazienza, Maria Teresa, Pennacchiotti, Marco & Zanzotto, Fabio Massimo (2005). Terminology extraction: An Analysis of Linguistic and Statistical Approaches. *Studies in Fuzziness and Soft Computing*, 185, 225–279. DOI: 10.1007/3-540-32394-5_20.

Pearson, Jennifer (1998). *Terms in Context*. Amsterdam: John Benjamins.

Sánchez Aquilino & Cantos Gómez, Pascual (1997). Predictability of Word Forms (Types), and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus. *International Journal of Corpus Linguistics*, 2 (2), 251–272. DOI: 10.1075/ijcl.2.2.06san.

Scott, Mike (2008). *WordSmith* Tools version 5. Liverpool: Lexical Analysis Software. Available at www.lexically.net/wordsmith.

Sinclair, John (1991). *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

Sinclair, John (2005). Corpus and Text: Basic Principles. In Wynne 2005 (see below). Available at ota.ox.ac.uk/documents/creating/dlc/chapter1.htm.

Sinclair, Stéfan, Rockwell, Geoffrey & the Voyant Tools team (2012). Voyant Tools [Computer software]. Retrieved from www.voyant-tools.org.

Sparck Jones, Karen (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11–21. DOI: 10.1108/eb026526.

Sternfeld, Joshua (2012). Pedagogical Principles of Digital historiography. In Hirsch (Ed.), *Digital Humanities Pedagogy*. London: Open Book Publishers. Available at books.openedition.org/obp/1645.

Stubbs, Michael (2001). *Words and Phrases*. London: Blackwell.

Tiersma, Peter (1999). *Legal Language*. Chicago: The University of Chicago Press.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: John-Benjamins.

Vivaldi, Jorge, Cabrera-Diego, Luis Adrián, Sierra, Gerardo & Pozzi, María (2012). Using Wikipedia to Validate the Terminology Found in a Corpus of Basic Textbooks. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*. Istanbul, Turkey. Retrieved from www.lrec-conf.org/proceedings/lrec2012/index.html.

Widdowson, Henry (2000). The limitations of linguistics applied. *Applied Linguistics*, 21 (1), 3–25. DOI: 10.1093/applin/21.1.3.

Williams, Geoffrey (1998). Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics*, 3(1), 151–171. DOI: 10.1075/ijcl.3.1.07wil.

Williams, Geoffrey (2001). Mediating between lexis and texts: collocational networks in specialised corpora. *ASp, la revue du GERAS*, 31, 63–76. Available at asp.revues.org/1782.

Wynne, Michael (Ed.) (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow books. Retrieved from ota.ox.ac.uk/documents/creating/dlc.

International Journal of
Language & Law

Breeze, JLL 6 (2017): 1–17
www.languageandlaw.de

Corpora and Computation in Teaching Law and Language

*Ruth Breeze**

Abstract
Corpus use has revolutionised the teaching of languages for specific purposes. In this article, I review some of the ways in which corpus studies can enhance our understanding of the regularities in legal language, looking particularly at formulaic language in four different areas of legal language: academic law articles, case law (judgments and opinions), documents (contracts, merger agreements, etc.) and legislation. After a brief overview of lexical issues in legal language, I look in greater detail at 4- to 8-word bundles in legal texts. After some consideration of legal modality and recurring syntactic structures, I show how these aspects come together with the phenomenon of bundles and formulaicity. I then provide some examples of how the kind of information provided here by specialised corpora can be exploited for teaching purposes.

Keywords
Legal English, corpus linguistics, multimodality, formulaic language, bundles, modal verbs

Submitted: 22 December 2016, accepted: 29 May 2017, published online: 28 June 2017

* *Ruth Breeze*: University of Navarra, Spain, rbreeze@unav.es. The author would like to thank the Instituto Cultura y Sociedad and the School of Law at the University of Navarra for their support.

DOI: 10.14762/jll.2017.001

1. Introduction

The construction of large corpora and the availability of increasingly sophisticated computer tools to process language information have profoundly changed the way we understand and teach languages.

On the one hand, it is now easier to search for words in context, both to gain a deeper understanding of the ways a particular word is used, its collocates, associated prosody, local grammar, etc., and to study lexical patterns that run through a range of material from different sources. On the other hand, we can also start from the text, and use corpus affordances to find out what makes a particular text or genre different from others, learn more about the contrast between spoken and written language, or gain insights into generic patterning that is not apparent to the naked eye. The insights from this not only help us to explain language phenomena more clearly to our students, but they also allow us to construct better didactic tasks and provide richer, more varied examples to use in the classroom.

All of this is true for language teaching in general, but it is even more important in the area of languages for specific purposes. Corpus use facilitates the creation of subject-specific wordlists, enhances the investigation of professional genres, and provides a wealth of insights into the socio-cultural phenomenon of specialised language. This article is intended to provide a selection of different ways in which corpora and computation can help us approach the teaching of legal English. Here, I describe how our understanding of aspects such as word frequency, keywords and bundles can be operationalised in preparing course material. This article considers the particular texture of legal texts in different genres, and the way in which formulaic language serves both to constitute the frames and fill the slots in legal discourse, particularly in the most highly conventionalised genres such as documents and legislation. My discussion then points to ways in which teachers can use corpora to gain deeper knowledge of complex text structures and formulaic expressions, in order to scaffold student learning.

2. Insights into specificity: why is legal English different?

Among specialised professional languages, legal English has the reputation of being one of the most difficult for the layperson. It presents challenges on many levels: in lexical areas (specialised terminology, often of Latin or Norman French origin, sometimes involving archaisms or redundant expressions), discourse organisation (very long sentences with many embedded clauses, the persistence of features such as compound reference words, such as “hereinafter”), interpersonality (performative speech acts, highly formal register, third-person address), grammar (frequency of conditional structures, characteristic modal system based on “shall” and “may”), and so on (Alcaraz

DOI: 10.14762/jll.2017.001

& Hughes, 2002). Despite initiatives such as the UK "Civil Procedure Rules" (1998) or the US "Plain Writing Act" (2010), much legal language remains inaccessible to non-specialists. Although studies based on close analysis of texts and interactions, or the diachronic development of particular genres, are valuable for understanding what is special about specialised language, computer-based investigation also offers a useful way of bringing out the unique nature of particular types of legal discourse. Corpus studies can potentially answer the question whether legal language is truly "different" from other kinds of English, that is, whether a "legal register" exists that runs through different genres, and what it might contain. They can also help us to show our students what variations occur from one legal genre to another, or even within one particular kind of text.

Legal English is, of course, a vast area containing many sub-domains which vary in terms of vocabulary, structures and genres. However, the huge expansion in international trade over the last twenty years has meant that the field of commercial law can be identified as particularly important for legal practitioners outside the English-speaking world. This means that law students taking degrees and LL.M.s are likely to benefit most from language support in this specific field. In order to approach this area of legal English using corpus linguistic tools, I gathered two million words from the area of commercial law, divided into four corpora of approximately 500,000 words each from: academic law articles on commercial and corporate law, case law (judgments and court opinions), legislation (Companies Acts) and legal documents (contracts, commercial lease agreements, merger agreements and so on) (see Breeze, 2013, for a more detailed description). WordSmith 6 and SketchEngine were used to perform the various quantitative tests used below.

Table 1: Comparative data in four legal corpora.

	Academic	Cases	Documents	Legislation
Lexical difference	3.69	3.55	5.69	5.73
Type/token ratio	38.52	35.52	29.43	24.16
Mean word length (in letters)	4.99	4.74	4.99	4.67
Mean sentence length (in words)	20.22	23.5	51.59	45.66

Note: *Lexical difference* is calculated by "compare corpus", using EnTenTen13 as a reference corpus. Higher numbers indicate greater differences.

Table 1 represents a starting point for the quantitative study of legal English. The table shows that the documents and legislation corpora differed more sharply from "general language" (represented by the EnTenTen13 corpus of general English) than the academic or case law corpora. Conversely, the type/token ratio was higher in academic and case law, and lower in legislation and documents, which is reasonable, given the technical nature of the lexis in legislation and documents: the same words are likely to be repeated for the sake of clarity, or because similar formulae are being used. The

mean sentence length was much greater in legislation and documents: The conventions governing these genres are quite different from those that characterise academic writing, since it is usual for a considerable amount of information to be included in one sentence, and it is possible for a single sentence to extend over several paragraphs or sections of the text.

3. Lexical issues

Like most professional areas, the law is rich in specialised terminology. Technical terms are an essential feature of specialised areas. Moreover, it is also likely that certain types of written document (instructions, technical reports, etc.) will rely more heavily on technical terminology than, say, promotional websites or letters to clients. As we saw in Table 1, although all four corpora presented a substantial degree of lexical difference from the general English reference corpus (EnTenTen13), the Documents and Legislation corpora contrasted more dramatically with the reference corpus, which indicates that the vocabulary of these corpora is much more specialised.

The measure of keyness allows us to find out which words are particularly frequent in the corpus in question, in comparison with the larger reference corpus (in this case, EnTenTen13). To show how this can be used, Table 2 displays the eight verbs in each corpus with the highest keyness scores.

Table 2: Verbs with highest keyness score (reference corpus: EnTenTen13) in each corpus.

Academic		**Cases**		**Documents**		**Legislation**	
Arbitrate	701	Dismiss	64	Contribute	304	Authorise	107
See	101	See	61	Indemnify	258	Allot	69
Litigate	65	Allege	54	Affiliate	159	Restate	59
Pre-empt	36	Abet	45	Assume	106	Contravene	55
Enforce	32	Imply	37	Exclude	105	Specify	52
Liquidate	29	Litigate	34	Contemplate	94	Confer	49
Preclude	29	Subrogate	32	Amend	87	Comply	45

As Table 2 shows, many verbs with a highly technical meaning have a high keyness score, which means that technical words are very frequent and so learners will need to be familiar with them in order to make progress in their comprehension of legal texts. Similar data could be presented for nouns or adjectives, shedding light on the need to emphasise technical lexis when teaching legal language. The use of corpora also makes it possible to zoom in on particular specialised areas within one field. So, for example, if we compare a "minicorpus" of contracts of sale with the main Documents corpus, we

can use the keywords function to find out what lexical items are going to be particularly frequent when we are looking at contracts of sale. As Table 3 illustrates, some of these words are predictable, while others might come as a surprise. When compiling course material, use of such procedures can help teachers to ensure that students have adequate vocabulary coverage from the type of document they are likely to encounter, or from a specific range of documents.

Table 3: Keywords in subcorpus of contracts of sale (keyness >45).

buyer	closing	allocated	preferential
defect	assets	intangible	records
seller	title	escrow	affected
purchase	knowledge	past	transaction

Information about the vocabulary that is specific to each area can be used, in combination with our knowledge of the lexis of legal texts in general, to generate practice and revision exercises such as Exercise One. Such exercises at first appear difficult, because of the clustering of unfamiliar vocabulary within a complex sentence. However, when students learn to approach the task systematically, they soon find that their understanding of the legal background and the interactional character of the clause enables them to solve the problem easily and build up confidence to tackle longer texts.

Exercise One

Put the words in bold into the correct gap in this extract from a contract of sale:

Escrow **Buyer** **Sellers** **Claims**

Promptly upon the expiration of the Claims Period, …… (*Answer:* Buyer) shall pay to …… (*Answer:* Sellers) an aggregate amount equal to that portion of the …… (*Answer:* Escrow) Amount that has not been used to satisfy Buyer's indemnification …… (*Answer:* Claims).

Nonetheless, since legal English textbooks often have a lexical orientation (cf. Krois-Lindner, 2006; Brown & Rice, 2007; Reinhart, 2007; see also Breeze, 2015), and students are generally extremely aware of the need to acquire a large specialised vocabulary, this will not form the main object of the present paper.

4. Exploring formulaic language

Moving on from simple word frequencies and keyness to lexical patterning, the first feature that strikes many people when they read certain types of legal text is its highly formulaic nature. To examine formulaicity, I worked with the concept of the "lexical bundle", first applied in Biber et al. (1999), which is specifically used to mean frequently recurring sequences of words regardless of the nature of the kind of links that might exist between them. In other words, such "bundles" may not be collocations or set

phrases, but the fact that they recur frequently may have some significance for our understanding of specialised language (Biber & Conrad, 1999). Biber (2006) brought to light patterns that emerge from seemingly fragmentary bundles, revealing a certain degree of regularity within fragmentation. For example, he documented the presence of large numbers of stance expression fragments, discourse organising fragments, and referential expressions, as well as a certain number of set phrases. Other authors working on spoken academic discourse have shown that such bundles often have discourse organising functions (Csomay, 2004; Nesi & Basturkmen, 2006). Studies of academic written language, on the other hand, have shown that bundle use varies across disciplines (Hyland, 2008): research-oriented bundles used in the sciences prioritised empirical methods and findings, while text-oriented bundles in humanities and social science disciplines reflected the value accorded to coherent argument (Hyland, 2008: 16). This section will show how examination of bundles sheds light on legal discourses and provides material that can be exploited pedagogically.

When the four corpora in this study were examined using WordSmith to identify frequent bundles, the corpus with the greatest number of repeated combinations of 4 to 8 words was the legislation corpus, followed by the documents corpus. The academic and cases corpora made use of fewer long bundles, although they did have more 4- and 5-word bundles than would be expected in, say, fiction or media texts.

Table 4: Frequency of different 4- to 8-word bundles in the four corpora (from Breeze, 2013).

	Academic	Cases	Documents	Legislation
8-word bundles	0	1	19	54
7-word bundles	1	3	38	75
6-word bundles	2	5	80	115
5-word bundles	8	18	171	284
4-word bundles	53	76	384	564

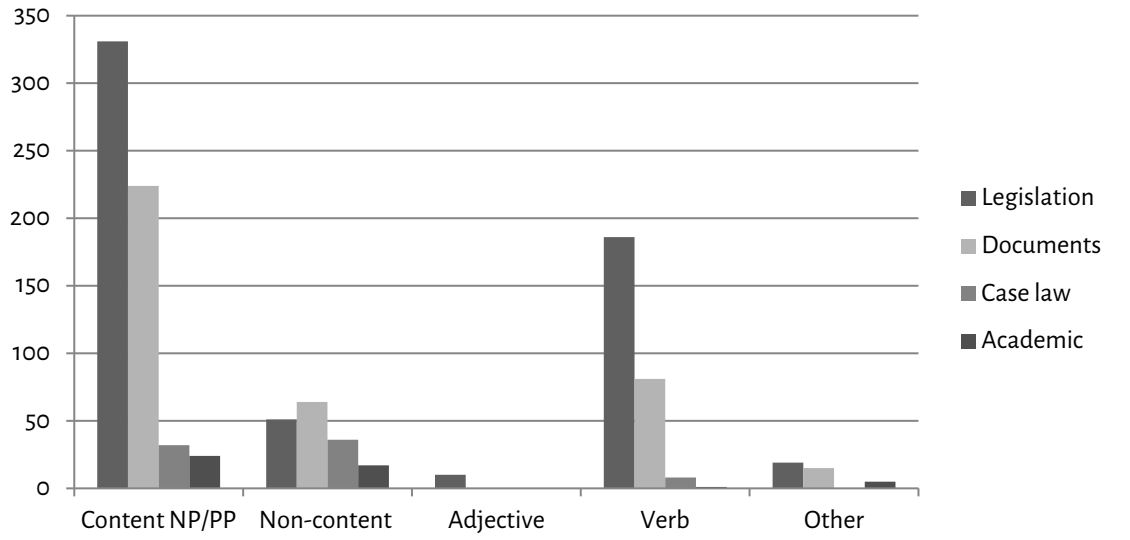
Note: Numbers denote raw (absolute) frequencies in each corpus of approximately 0.5 million words.

As Table 4 suggests, the most striking bundles were those of 5, 6, 7 and 8 words found in the Documents and Legislation corpora. These tended to be either heavy noun phrases such as "officer or secretary of the board of directors", or verb phrases such as "shall have the meaning set forth in the" which reflect formulae used in documents. Although such phrases might not seem particularly attractive pedagogically, we should bear in mind that speed of comprehension (as well as production) generally improves when students learn to recognize (or produce) fixed or semi-fixed lexical chunks (Nattinger & DeCarrico, 1992; Wray, 2000). Since many legal documents consist of sentences like example 1 (below), the ability to recognize and process fixed formulae is a skill that students should acquire. Focusing students' attention on how to divide the sentence into its component chunks is likely to be useful for comprehension purposes, and absolutely essential if translation forms part of the curriculum.

(1) No waiver by either party hereto / of any breach of this Agreement / shall be deemed to be / a waiver / of any preceding or succeeding breach / of the same or any other provision hereof.

Following on with the topic of bundles, since 4-word bundles were frequent, we focused on classifying these, using a procedure based on Biber (2006). Four main categories emerged: content noun phrases, non-content phrases, verb phrases, and instructions. Around 4% of the 4-word-bundles had to be discarded because it was not clear which category they might belong to.

Figure 1: Bundle types in the four corpora.



From the information displayed in Figure 1, it is evident that the category of "non-content" bundles accounted for a significant proportion of these bundles. On closer inspection, many of these turn out to be complex prepositional phrases such as "in the context of", "on behalf of the", "at the time of" or "in the event of". As in the case of the longer bundles, familiarity with these patterns should help students to gain reading speed and improve their accuracy.

Table 5: Ten most frequent 4-word prepositional phrases in Documents and Legislation corpora.

Rank	Documents	Legislation
1.	In accordance with the	In the case of
2.	On behalf of the	For the purpose of
3.	With respect to the	In accordance with the
4.	In connection with the	In respect of the
5.	In the case of	With respect to the
6.	In respect of the	On behalf of the
7.	As a result of	Within the meaning of
8.	To the extent that	To the extent that
9.	To the knowledge of	As a result of
10.	In the event of	In the event of

Since 4-word bundles are particularly frequent in the Documents and Legislation corpora, Table 5 shows the ones which occur most in each corpus. The most common prepositional phrase bundles appear in the context of the need for inter- and intratextual reference in legal texts (i.e. "in accordance with the"), the need for delimitation and precision (i.e. "to the extent that"), and a process of nominalisation of causal and conditional relations (i.e. "as a result of" to replace "because", and "in the event of" to replace "if"), which has been discussed elsewhere as a typical feature arising from the need to assign technical legal values to actions or utterances (Vázquez Orta, 2010: 273). In this context, exercises of the following type can be used to raise students' awareness of this type of bundle.

Exercise Two

Complete these phrases from legal documents and laws using "in", "on", "within" or "of":

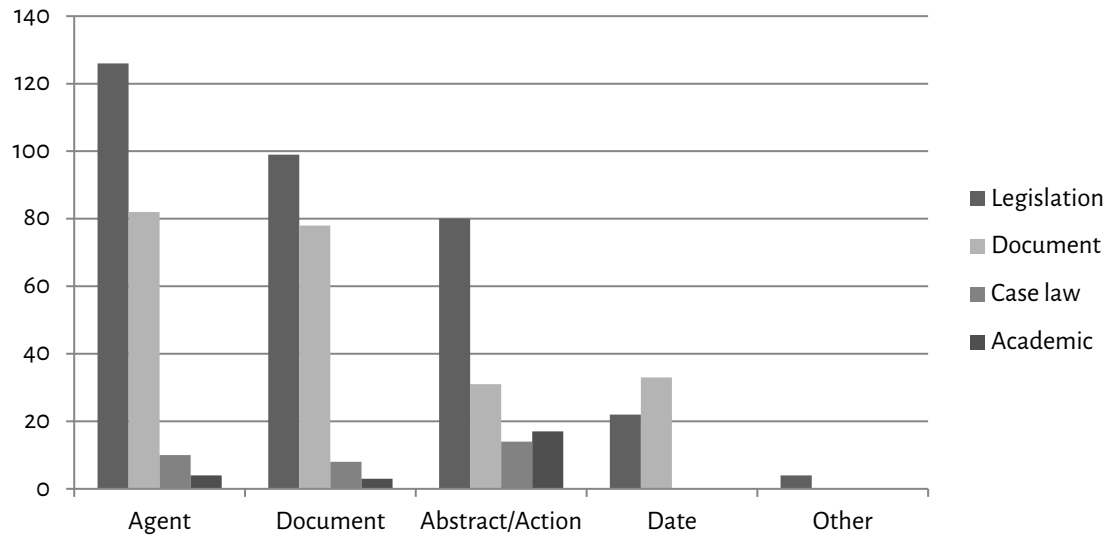
A. (…… *Answer:* In) the event (…… *Answer:* of) a breach or threatened breach of the terms of this Agreement by Consultant, the parties hereto agree that monetary damages would be an inadequate remedy for said breach.

B. Any person who is a worker (…… *Answer:* in) the meaning (…… *Answer:* of) the Act and is over 16 years of age may join a trade union.

C. When an arbitrator considers that a statement of claim made (…… *Answer:* on) behalf (…… *Answer:* of) the claimant should be the subject of two or more separate arbitrations, he may refuse to deal with multiple claims in a single reference.

Since content bundles are frequent, it is also interesting to examine which type of content bundle is more frequent in the different corpora. Figure 2 below shows the proportion of 4-word bundles belonging to the category of "content" that could be classified as representing agents (people or institutions), documents (laws, contracts, etc.), and abstract concepts or actions.

Figure 2: Bundle categories in the four corpora.



As we can see from Figure 2, the names of documents and abstract concepts figured largely in these texts. Documents ranged from legislation ("Model Business Incorporation Act") to everyday documents in the life of a company ("this memorandum of association"). Abstracts ranged from theoretical entities such as "contractual choice of law" or "the corporate law market" in academic texts, to aspects of corporate practice ("ordinary course of business", "all liabilities and obligations") in the documents corpus. Again, the frequent bundles can be identified and used in the classroom in order to familiarize students with the texture of legal texts.

Exercise Three provides a scaffolded approach to understanding a dense clause concerning the bundle "a Material Adverse Effect", embedded within a complex grammatical structure. The twofold difficulty (heavy noun phrase and conditional passive of "expect") is disentangled stage by stage, as students are invited to try to express the same legal concept in everyday language.

Exercise Three

Read the following clause from a merger agreement. You are going to explain this clause to a client who is not a legal specialist. Make some notes to help you give your explanation.

Absence of Certain Changes. Since December 31, 2007 until the date hereof, (1) the Company and the Company Subsidiaries have conducted their respective businesses in all material respects in the ordinary course, consistent with prior practice, (2) except for publicly disclosed ordinary dividends on the Common Stock and outstanding Company Preferred Stock, the Company has not made or declared any distribution in cash or in kind to its shareholders or issued or repurchased any shares of its capital stock or other equity interests and (3) no event or events have occurred that has had or would reasonably be expected to have a Material Adverse Effect.

Before you give your explanation, answer the following comprehension questions:

1. What is a Material Adverse Effect? (*Answer:* Something that has happened in the company that would make it less attractive to buy.)

2. Re-read the final phrase: "no event or events have occurred that has had or would reasonably be expected to have a Material Adverse Effect". Try to express this without using the passive. (*Answer:* Nothing has happened in the company that would make people think that it is less attractive to buy.)

3. If you have problems with the last sentence, check the way the passive is used here. We can say that "we expect that an event will have a Material Adverse Effect". How can we express this idea impersonally, using "is expected to"? (*Answer:* An event is expected to have an MAE.) How can we express this to make it sound as if we are not certain? (*Answer:* An event could be expected to have an MAE.) How can we express this to make it sound as if people would reasonably expect the event to have an MAE? (*Answer:* An event would be reasonably expected to have an MAE.)

Now work in pairs, taking turns to be the client who does not understand the clause and lawyer who has to explain the clause.

5. Grammatical patterns

As the examples in the previous section show, the degree of grammatical complexity in legal texts is often very high, and this phenomenon presents a considerable challenge

for teachers. One approach to this might be to identify some of the reasons why the sentences in Documents and Legislation (see Table 1) are so long, in terms of the pragmatic functions that are being fulfilled and the conventions associated with their realisation. One particular function that has been associated with legal texts since Babylonian times (see Breeze, 2013) is that of seeking to define the consequences of actions, or the conditions in which something may or must be done. The use of "if" is frequent in this context, even though other mechanisms also exist (such as "in the case of" or "in the event of" used with nominalisations, as mentioned above).

Table 6: Frequency per 10,000 words of "if" and "if"/"had" in the four corpora.

	Academic	Cases	Documents	Legislation
If	23.1	44.0	25.4	43.9
If + (1-4) + had	0.5	1.3	0.3	1.3

Table 6 shows the frequency of "if" (as a crude measure of conditional structures in which verbs are used) and "if" plus "had" (within five words either way, calculated using the "filter" option on SketchEngine) (as a crude measure of counterfactual conditionals) in the four corpora. Interestingly, although the frequency of "if" in the Academic and Documents corpus is similar to that found in BNC (22 per 10,000) and EnTenTen13 (24.6), the frequency in Cases and Legislation is much higher. This is logical, in that conditional-type structures are strongly associated with proceedings and texts in which different courses of action and their consequences are laid out. It is slightly more difficult to see why such structures are less frequent in Documents, but the answer would appear to lie in the point that such documents (mainly contracts of different kinds) exist to set out what must and must not be done in a particular situation, rather than to allow for many different contingencies (as in the case of legislation) or to determine whether or not something was actually done, and whether or not that action violated a particular rule or principle (as in case law).

The following simple exercise based on an extract from the Legislation corpus illustrates the way students can be encouraged to develop an awareness of the characteristic "if" structures in legal texts. When used with law students, this type of exercise serves to draw students' attention to specific features of the language of the text, as well as to encourage close reading. It is likely that the same type of exercise would also have potential for use with, say, students of translation, but in this case, the analysis could be directed towards linguistic aspects, such as the difference between "limited by" and "limited to", or the reasons motivating the use of repetitions in legal genres.

Exercise Four

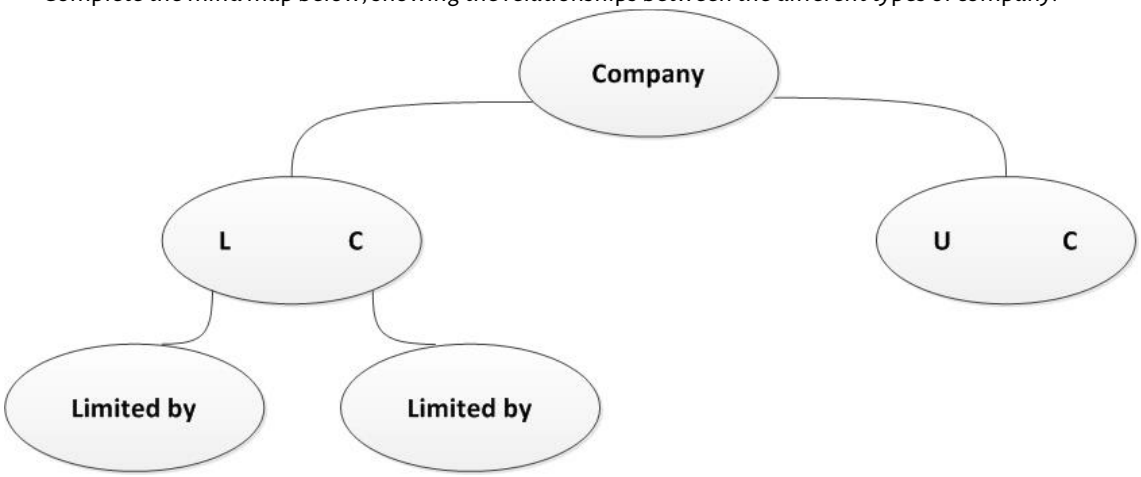
Read the following extract about different types of company:

A company is a "limited company" if the liability of its members is limited by its constitution. It may be limited by shares or limited by guarantee. If their liability is limited to the amount, if any, unpaid on the shares held by them, the company is

"limited by shares". If their liability is limited to such amount as the members undertake to contribute to the assets of the company in the event of its being wound up, the company is "limited by guarantee". If there is no limit on the liability of its members, the company is an "unlimited company".

The word "if" occurs five times in the above extract, and one of the uses is different from all the others. Which one is different?

Complete the mind map below, showing the relationships between the different types of company.



6. Modality

It is well known that the system of modality in English-language legal documentation traditionally finds its central axis in "shall" (obligation) and "may" (permission) (Trosborg, 1997). Here, we find that "shall" is the 9th most frequent word in the Documents corpus, and is also common in legislation, while in the Academic and Case law corpora the frequency of "shall" is much closer to that found in general English. On the other hand, "may" is particularly frequent in Legislation (seemingly, legislators are more interested in granting permission than in prescribing). All four legal corpora have higher frequencies of "may" than the general corpus (see Breeze, 2014, for an investigation of the use of "may" in legal correspondence). Regarding the other modal verbs quantified here, "will" is much more frequent in the general corpus than in the legal corpora, while "should" occurs more in the Academic and Case law corpora, and hardly at all in Documents and Legislation (which are not associated with discourse functions such as advising, recommending or predicting commonly fulfilled by "should").

Table 7: Frequency per 10,000 words of modal verbs of obligation and permission across corpora.

	Academic	Cases	Documents	Legislation	EnTenTen13
Shall	3.2	4.4	103.5	39.6	1.0
May	27.1	13.4	28.7	50	11.0
Will	19.8	10.9	17.8	28.9	38.8

	Academic	Cases	Documents	Legislation	EnTenTen13
Must	7.4	8.8	2.5	29.9	4.7
Should	12.8	11.8	1.2	1.4	8.9

If we combine the information we have about the frequency of "shall" with what we know about verb phrase bundles (see above), we find that "shall" commonly occurs in bundles such as "shall be read as" and "shall be deemed to" in Legislation, and "shall be governed by", "shall have the right", and "shall be deemed to" in Documents. Although the use of "must" rather than "shall" is preferred by some advocates of plain English (Garner, 2001), change has been slow outside the United States, particularly among drafters of legislation and legal documents (Williams, 2011): for example, Williams (2013) shows that the frequency of "shall" remained stable among EU drafters between 1973 and 2010. In fact, although use of "shall" for obligation is an archaism, its use in technical legal contexts rarely presents difficulties of comprehension once the reader is aware of this convention. Exercise Five below is designed to encourage students to focus on the ways in which one frequent bundle is used in legal documents. Such exercises should enable law students to become more familiar with the language of documents and learn to take advantage of its formulaic nature in order to read more efficiently. Exercises in chunking are also likely to bring benefits for translation students.

Exercise Five

In legal documents, the technical phrase "shall be deemed to" has a special role in spelling out the terms of an agreement or understanding in order to ensure shared comprehension. Look at the following extracts from legal documents, and complete the table below. Then answer questions A to C.

1. *If any notice of a proposed sale of Guaranty Collateral shall be required by law, such notice shall be deemed reasonable and proper if given at least ten (10) days before such sale or other Disposition.*
2. *If any term or provision of this Guaranty shall be deemed prohibited by or invalid under any applicable law, such provision shall be invalidated without affecting the remaining provisions of this Guaranty.*
3. *Nothing in this Agreement shall be deemed to create or constitute a partnership, agency, representative or other relationship between the Parties.*
4. *No failure on the part of a Party to exercise or delay in exercising any right hereunder will be deemed a waiver thereof, nor will any single or partial exercise preclude any further or other exercise of such or any other right.*
5. *Any notice, request, instruction or other document to be given hereunder by any party to the other will be in writing and will be deemed to have been duly given (a) on the date of delivery if delivered personally or by telecopy or facsimile, upon confirmation of receipt, (b) on the first business day following the date of dispatch if delivered by a recognized next-day courier service, or (c) on the third business day following the date of mailing if delivered by registered or certified mail, return receipt requested, postage prepaid.*

Now fill in the table below about how the verb "to deem" is used in legal documents, using examples from the box above:

With infinitive: (*Answer:* shall be deemed to create)
With past infinitive: (*Answer:* will be deemed to have been duly given)
With adjective: (*Answer:* shall be deemed reasonable and proper)
With past participle: (*Answer:* shall be deemed prohibited)
With noun: (*Answer:* will be deemed a waiver)
Read the instructions and complete the sentences in each case:

A. You want to define the term "employee" so that it excludes members of the board of directors. Use "shall be deemed" plus infinitive to complete the sentence from the articles of association.

For purposes of this Article III (and the definition of "Maximum Option Number") only, the term "Employee" …….. (*Answer:* shall be deemed to exclude) members of the Board of Directors of the Company.

B. You need to say that nothing in this agreement is supposed to create third party rights.

Nothing in this Agreement …….. (*Answer:* shall be deemed to create) any third party beneficiary rights in any Person or entity not a party to this Agreement.

C. You need to say that stockholders who do not make a proper stock election in the correct way will be understood not to have made a stock election.

Any holder of Company Common Stock who does not properly make a Common Stock Election prior to 5:00 p.m., New York City time, on the Election Date, …….. (*Answer:* shall be deemed not to have made / shall not be deemed to have made) a Common Stock Election, and all of such holder's Company Common Shares shall be converted into the right to receive the cash merger consideration as set forth in Section 3.2(c)(i), subject to Section 3.2(c)(iv).

7. Verb-preposition combinations

Given the difficulty of legal English for students, it is perhaps interesting to look briefly at an area where legal language may actually present learners with fewer problems than general English does. Studies of legal English have generally paid little attention to verb/preposition combinations, and where textbooks have done so, the focus is usually on verbs with dependent prepositions (e.g. Krois-Lindner, 2006, contains many useful exercises involving typical combinations of verb and preposition), although phrasal verbs are also found (e.g. carry out, spin off). Table 8 shows the main verb-plus-preposition combinations found in the four corpora, obtained using corpus query language in SketchEngine.

Table 8: Ten most frequent verb-preposition combinations in the four corpora (raw frequencies).

Rank	Academic		Cases		Documents		Legislation	
1.	Base on	188	Deal with	51	Comply with	119	Comply with	197
2.	Apply to	141	Apply to	35	Enter into	73	Apply to	195
3.	Relate to	115	Enter into	32	Result in	51	Deliver to	56
4.	Enter into	115	Engage in	27	Deliver to	51	Vote on	47
5.	Depend on	90	Dispose of	27	Apply to	44	Provide for	36
6.	Refer to	69	Comply with	27	Participate in	37	Subscribe for	26
7.	Assign to	63	Result in	25	Pay to	26	Participate in	25
8.	Provide for	60	Rely on	24	Inure to	25	Dispose of	22
9.	Deal with	51	Refer to	22	Cooperate with	24	Carry on	22
10.	Involve in	48	Account for	21	Interfere with	23	Apply for	18

Knowledge of the verb-preposition combinations that students are most likely to encounter in different genres enables teachers to compile exercises that focus attention on particular issues that are likely to be problematic, such as the choice of preposition. Exercise Six was built from the Documents corpus, and represents a model for centering students' attention on frequent verb-preposition combinations. Such exercises allow students to practise the notoriously difficult area of dependent prepositions in a specifically legal context.

Exercise Six

Choose an appropriate preposition from the list in bold to fill the gaps in the following extracts from legal documents:

with **to** **in** **into** **with**

1. Such Investor has the corporate or other power and authority to enter …….. (*Answer:* into) this Agreement.
2. Tenant shall comply …….. (*Answer:* with) any reasonable regulations made by the Landlord regarding the use and occupation of the Premises.
3. The Parties shall not do anything that might interfere …….. (*Answer:* with), obstruct or delay the satisfaction of all or any of the Conditions.
4. The provisions of this Agreement shall be binding upon and inure …….. (*Answer:* to) the benefit of the Parties and their respective successors and permitted assigns.
5. The Company shall use its reasonable efforts to structure the Transactions in a manner that does not result …….. (*Answer:* in) any material tax to the Executive.

8. Textures and trends

Returning to our general overview of formulaic language and typical patterns above, it is possible to put together some ideas about how bundles and syntactic patterning work in the different corpora, and therefore in the different genres. Of course, all language may be underpinned by a restricted range of structural possibilities which offer slots to a vast range of lexical options in order to generate an almost infinite series of meanings. However, what makes specialised language particularly fascinating is that the structures, slot-fillers and meanings are all much more closely circumscribed.

Legal language offers a range of "textures" in this sense, going from the more varied, more loosely structured organisation of academic discourse, to the highly formulaic, tightly structured language of documents and legislation. In what follows, I outline the main findings from the four corpora in terms of the way formulaic language works in each, thinking particularly of the different functions which bundles appear to have in the different types of text.

Academic legal texts contain abstract conceptual noun phrases which act as placeholders, but also non-content prepositional phrases used for referential framing:

(2) The illogic in preserving a distinction between *void and voidable contracts* can be illustrated *in the context of* a contract that is void ab initio for illegality.

Case law contains many verb phrases representing speech acts or indicating textual orientation. Place-holders are often noun phrases, which can be concrete (actors, documents) or abstract (concepts). Non-content prepositional phrases are used for referential framing:

(3) The Claimant was then to divide up the money *in accordance with the* other *terms of the partnership* to which I shall refer below.

Documents contain many heavy noun phrases (actors, documents, concepts) as place-holders (4), and many non-content prepositional phrases for referential framing, as well as post-modifiers that are also used for intra- and intertextual reference (5):

(4) A shareholder may also take action against *another shareholder or director* pursuant to *these Articles of Association*.

(5) If such disclosure is made *in accordance with the* confidentiality obligations *set forth in this* Agreement.

Legislation abounds in prepositional phrases which orient the reader within the text or towards other documents, as well as performing functions related to referential framing. It also contains many deontic/regulatory phrases used to connect concepts together. On the other hand, heavy noun phrases (actors, documents and concepts) act as place-holders.

(6) Nothing in the preceding provisions of this section / shall be construed as preventing the use of a registered trademark by any person for the purpose of identifying goods and services.

If we put these ideas together in terms of frames (discourse structures) and slots (places for noun phrases, such as actors, documents or concepts), extract (7), from the articles of association of a company, can be chunked in various ways, and could be presented either in the form of a frame awaiting slots (8), or in the form of place-holders to be joined together by a frame (9).

(7) The board of directors' resolutions / in respect of / all other matters / may be passed by / the affirmative vote of / a simple majority of the directors.

(8) ..... in respect of ..... may be passed by .....

(9) The board of directors' resolutions ..... all other matters ..... the affirmative vote of / a simple majority of the directors.

Similarly, the legislation clause below (10) could be divided as proposed here, and then presented as a frame (11) or as place-holders (12).

(10) A person guilty of an offence / is liable on summary conviction to a fine / not exceeding level 3 / on the standard scale / and, / for continued contravention, / a daily default fine / not exceeding one tenth of level 3 / on the standard scale.

(11) ..... is liable on summary conviction to ..... and, for continued contravention, ...

(12) A person guilty of an offence ..... a fine not exceeding level 3 on the standard scale ..... a daily default fine not exceeding one tenth of level 3 on the standard scale.

As teachers, it is extremely important for us to show students how this type of slot-frame interaction occurs in the texts we use. With some support, law students will be able to draw on their familiarity with the interactional framework of legal discourse to decode the text satisfactorily. For translation students, on the other hand, guided work with structures of this kind will help them build an awareness of the law as a system, with its actors, actions, eventualities and consequences, and to gain a feeling for the special pragmatics which underpins the language of the law.

9. Concluding reflections

For people who teach legal English, it is essential to make students aware not only of specific high-frequency terminology, but also of the typical formulaic language that they will encounter. This can be done by consciousness-raising exercises, and by setting tasks such as those exemplified in this article. On a basic level, students should then also be encouraged to develop chunking skills, so that they can read and interpret legal documents or legislation more easily. On a more advanced level, it is also important for student to gain hands-on user knowledge of the typical framework structures that sustain particular legal genres, particularly different types of contract clause. For example, students can be given model clauses or templates that have to be completed using specific information, or clauses that have to be corrected or adapted to new situations.

Looking ahead, we need further research based on larger and wider corpora in order to examine how the type of formulaic language identified here behaves in other legal genres and domains. Moreover, since our current knowledge is based mainly on written evidence, it would be stimulating to examine how formulaic language operates in multimodal terms, looking at spoken legal language across a range of contexts. This would enable us to overcome the distortions imposed by the exclusive focus on textual evidence, and develop a more ecologically valid understanding of legal language as a spoken system with pragmatic, interactional and even performative dimensions. We also need to work on how to exploit corpora in the classroom, with consideration of how the kind of information provided by corpora can be accessed and used productively by students themselves. New technological affordances, such as searchable learner-friendly corpora or multimodal corpora, are currently opening up exciting new possibilities for research and teaching in this area.

References

Alcaraz, Enrique & Hughes, Brian (2002). *Legal translation explained*. Manchester: St Jerome Publishing.

Biber, Douglas (2006). *University language: a corpus-based study*. Amsterdam: John Benjamins.

Biber, Douglas & Conrad, Susan (1999). Lexical bundles in conversation and academic prose. In Hasselgård & Oksefell (Eds.), *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 181–189). Amsterdam: Rodopi.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Breeze, Ruth (2013). Lexical bundles in four legal genres. *International Journal of Corpus Linguistics, 18*(2), 229–253. DOI: 10.1075/ijcl.18.2.03bre.

Breeze, Ruth (2014). The discursive construction of professional relationships through the legal letter of advice. In Breeze et al. (Eds.), *Interpersonality in legal genres* (pp. 281–302). Bern: Peter Lang.

Breeze, Ruth (2015). Teaching the vocabulary of legal documents: a corpus-driven approach. *ESP Today 3*(1), 44–63. Available at www.esptodayjournal.org/esp_today_back_issues_vol4.html.

Brown, Gillian & Rice, Sally (2007). *Professional English in Use: Law*. Cambridge: Cambridge University Press.

Csomay, Eniko (2004). Linguistic variation within university classroom talk: A corpus-based perspective. *Linguistics and Education, 15*(3), 243–274. DOI: 10.1016/j.linged.2005.03.001.

Garner, Brian (2001). *Legal writing in plain English* (2nd ed.). Chicago: University of Chicago Press.

Hyland, Ken (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21. DOI: 10.1016/j.esp.2007.06.001.

Krois-Lindner, Amy (2006). *International Legal English*. Cambridge: Cambridge University Press.

Nattinger, James & DeCarrico, Jeanette (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesi, Hilary & Basturkmen, Helen (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283–304. DOI: 10.1075/bct.17.03nes.

Reinhart, Susan (2007). *Strategies for legal case reading and vocabulary development*. Ann Arbor: University of Michigan Press.

Trosborg, Anna (1997). Text Typology: Register, Genre and Text Type. In Trosborg (Ed.), *Text Typology and Translation* (pp. 3–23). Amsterdam/Philadelphia: Benjamins. DOI: 10.1075/btl.26.03tro.

Vázquez Orta, Ignacio (2010). A genre-based view of judgments of appellate courts in the common law system. In Gotti & Williams (Eds.), *Legal discourse across languages and cultures* (pp. 263–284). Bern: Peter Lang.

Williams, Christopher (2011). Legal English and Plain language: an update. *ESP Across Cultures, 8*, 139–151.

Williams, Christopher (2013). Changes in the verb phrase in legislative language in English. In Aarts et al. (Eds.), *The verb phrase in English: Investigating recent language change with corpora* (pp. 353–371). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139060998.015.

Wray, Alison (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics, 21*(4), 463–489. DOI: 10.1093/applin/21.4.463.

International Journal of
Language & Law

Vogel et al., JLL 6 (2017): 90–100
www.languageandlaw.de

"Begin at the beginning"
— Lawyers and Linguists Together in Wonderland

Friedemann Vogel, Hanjo Hamann, Dieter Stein,
*Andreas Abegg, Łucja Biel, and Lawrence M. Solan**

Abstract
What do patterns in legal language tell us about power, policy and justice? This question was at the heart of a conference on "The Fabric of Language and Law: Discovering Patterns through Legal Corpus Linguistics", convened in March 2016 by the international research group "Computer Assisted Legal Linguistics" (CAL²) under the auspices of the Heidelberg Academy of Sciences. About forty scholars from Germany, Switzerland, Italy, Poland, Spain and the US brought together their different intellectual and disciplinary perspectives on computational linguistics and legal thinking. Concluding the conference, four legal linguistics experts – two native linguists, two native lawyers – discussed the perspectives and limitations of computer-assisted legal linguistics. Their debate, which this article faithfully reproduces, touches on some of the essential epistemological issues of interdisciplinary research and evidence-based policy, and marks the way forward for legal corpus linguistics.

Keywords
corpus linguistics, law and language, legal linguistics, fabric, pattern, CAL², panel discussion

Editorial (not reviewed), first published in The Winnower 2016, republished in JLL: 7 September 2017

* *Vogel:* Institute of Media Culture Science, University of Freiburg, Germany, friedemann.vogel@mkw.uni-freiburg.de; *Hamann*: Max Planck Institute for Research on Collective Goods, Bonn, Germany, hamann@coll.mpg.de; *Stein:* English Department, Heinrich Heine University, Düsseldorf, Germany, *Abegg*: Center for Public Commercial Law, ZHAW School of Management and Law, Winterthur, Switzerland; *Biel*: Institute of Applied Linguistics, University of Warsaw, Poland; *Solan*: Brooklyn Law School, Brooklyn NY, USA.

The State of the Art in Legal Corpus Linguistics

Faithful transcript of a panel discussion on 19th March 2016 in Heidelberg, chaired by a co-founder of the International Language and Law Association (ILLA), Dieter Stein. The text was edited sparingly for legibility, and typographically emphasizes Stein's *moderation remarks* to distinguish them from his debate contributions. The transcript is identical with its prior publication in the open access journal *The Winnower* (DOI: 10.15200/winn.148184.43176) and was republished here merely for reference.

Dieter Stein: *What had you not gotten to know, had you not come here for this meeting?*

Łucja Biel: I learned a lot about the fabric of law as such. I really like the concept of the fabric, because I think it nicely combines with the way we think of texts, the new way we think of texts, the new way we perceive texts.

I also learned that there are different views as to what a pattern is. This is very meaningful because there are different types of patterns and in particular you can see differences between people who work in different disciplines. For example lawyers, like Larry Solan, work at a very high conceptual level, with the understanding of a pattern as a rule. Perhaps lawyers will be more interested in the conceptual structure and how it is patterned. Linguists and people who work with language acquisition, like Ruth or María, had a somewhat different conception of a pattern, working with n-grams for example, trying to extract multiword terms. It was also very interesting to see that computational linguists, like Giulia, can think of patterns as the depth of embedding and the complexity of embedding. So you have all these very different perspectives and we should think of some common ground to integrate all those views and all those levels of patterns, at different levels of language organisation.

That leads me to a question: Is there any universality of such patterns between languages? Because it was very useful to see how we work with different languages, and what kind of patterns we have in different languages, and I think there is some common ground between the languages. We have to deal with how to compare patterns between languages so that we make those comparisons meaningful and methodologically balanced. Each corpus has a different composition, a different structure and design objectives, and this also makes it difficult to a certain extent to compare between languages.

I also learned that we have the growing availability of corpora, the growing resources. If you read any literature on the use of corpora and legal language or in particular legal translation, the first thing you learn is that it is so difficult to work with corpora, because there are scarcely any resources. Now I learned that we have a lot of resources right now, and perhaps, we have to communicate to the people who might be interested in working with those resources, who might have good ideas how to use those resources.

It was also very interesting to see how people with different academic interests approach corpora and what they do with the data, to see how corpora can be used.

Larry Solan: One thing I noticed is: The first two speakers were Americans and we immediately started talking about Big Data and cases. 'Here is what happened in this case, here is what happened in that case.' And then everybody else is working in a civil law environment, and almost everybody immediately started talking about legislation. Of course, the Americans and the Brits and the Canadians and the Australians have plenty of legislation also, we do not even have much common law left in the United States. Almost everything's statutory, not everything of course, but almost everything. And yet, we orient ourselves around the judges as opposed to orienting ourselves around the parliament and the congress.

Getting back to our question: At the end of the last paper, when the question was, how universal is this, are the German judges, and who are they going to quote? The answer is: They are going to cite the German legal literature, in German, that is what they are interested in. Picking up on your comments, there really is a kind of duality here, there are certain tools that are available universally, depending upon what language you are trying to work in. Those are the data. Then the corpus tools seem to be pretty well developed, I mean, everybody who wanted to do something technically, came here and talked about what they did, and more or less accomplished what they wanted to, in terms of organising corpora. There are difficulties, and there are soft spots, and Google is a terrific thing to criticize for just the right reasons, but generally speaking, there seems to me to be a great deal of success: In the kinds of tools, in the kinds of analyses people do – with some level of statistical sophistication, which probably should be higher –, and then the data are growing. It would be great to have just a resource bank where everybody can know where everything is, it probably could be collected in a couple of months, if a grad student wanted to do that.

So all of that is good, but then what use you put it to, some of that is universal, you could talk about information that reveals the hidden underbelly of comported rule of law values generally. But my guess is, it is not going to really work that way. My guess is that individual legal systems will be using it, using these tools towards either internal advances from within the system, such as in the United States deciding cases in the way Stephen and I were talking about, or improving legislation, as others were talking about when investigating the relationship between the supranational system and the individual countries. This is very important because Europe is so concerned about such things. You actually gave somebody some good news, we do not hear that too much these days.

Then there are many other tools that could be used in a domain specific manner. I really doubt, other than intellectuals in international law or comparative law, that there is going to be an enormous interest in something that is specifically about Swiss legislation. If you decided to give a talk in the United States, there is a group of international comparative law people, they come to it, they find it fascinating. Similarly, without universalising the problems, people always like to see pathologies in the United States system of justice because it feels good. But generally speaking, I think the

usefulness of these tools are likely, from what I have learned today, to be more domain specific. The tools themselves look like there is a big sophisticated international community of people who just know a lot about this stuff. That was really revealing and quite exciting for me!

Andreas Abegg: It was indeed very interesting to see the different projects and the creative ideas on empirical linguistics and law. With this very new method of empirical linguistics at hand, it is of great value to exchange on possible fields of application, to exchange on what approaches work (or do not work). It might be fruitful to establish a network or a site to collect and share ideas and to know about the current and finished projects. Such a network or site might also help to enter into new collaborations.

Furthermore, the different approaches by scholars from common law and continental law were of great interest to me. Common law lawyers do not find it difficult to immediately connect an empirical linguistic method to their case law. However, as continental law scholars, we cannot just concentrate on case law, but we always have to connect to legal principles which guide continental law. From our history, our path-dependency, we are much more into scholastic deduction. This makes it more difficult to use an empirical method. Therefore, because we do not have this immediate access to empiricism, there is a need for continental scholars to work on a theory that links empirical linguistics to the legal methods.

Dieter Stein: Maybe I will myself provide one or two remarks: For me, this is still kind of a new field, and if you have a new field, you have a situation where things are meandering a little bit until they really fall into place. I believe, this is the time now to establish some sort of a meta-theory of what we are doing.

You see, we have a bottom-up aspect here, we have many people, having wonderful work on corpora. But then arises the issue: What are we doing with this corpora? So what comes first? Do we construct the corpora first or do we first ask our questions and then construct the corpora? This is kind of a top-down perspective. And I would like to see a matching of those two perspectives. That is what I believe may still be something that we need to work on.

The second aspect to me is: Of course I was intrigued by the way legal language – language of law – just does not exist. You have a number of legal genres that are pretty much separate and these appear to me to be separate also in different countries. I was much intrigued by the work on evaluative elements in judgements by Stanisław. I would imagine this is not the same in all countries. What I would like to see is the theoretical instrument of genre, sharpened and applied to the analysis of legal language.

DOI: 10.14762/jll.2017.090

Open Debate: The Future of Legal Corpus Linguistics

Dieter Stein: *Let me now open the floor for discussion: Everyone can discuss, everyone can chip in, the audience is invited to comment.*

Larry Solan: I would like to respond to Andreas first and to you, Dieter. [...] I agree with both of you: This really requires nothing other than collaboration. I was teaching at a university, a few years ago, as visiting professor. Maybe ten years ago, they got their fMRI machine, somebody gave millions of dollars and they get and haul this thing in. It is very expensive to keep up and everything. They had no idea what to do with it. So they put signs up, 'Anybody have an experiment that you want to do with brain-imaging? That's great!', and then some people would sign up. Now it is really pretty sophisticated. One thing that the field is crying out for, is a collaboration between the identification of issues in law.

They could be very practical issues: How can we draft statutes that you can read? In the United States that is not much of a concern, we do not care whether you can read them. We care about whether they are precise and that there is not going to be ambiguities, but we do not care whether they are comprehensible. That is what lawyers are for, to spend their time reading them.

Or they could be at a high level of theory, they could be quite abstract, but it seems to me that most of the research is in the service of improving various aspects of the legal system. It could start with basic research, it does not have to have practical ramifications for the first generations. There is nothing wrong with that. When research funders require immediate gratification through practical consequences, that is a bad thing sometimes because it stifles basic research values. But it seems to me that this is really a direction.

Now, to the extent that this is work that just happens to be in the law – because language for special purposes and corpus research generally is something that people are interested in and law is just a nice domain to do it in, because there is learning within the linguistics – nothing what I am saying really applies. But to the extent that linguists sometimes are frustrated by the lack of attention they get from the legal community, it is not easy to start with a perspective that the target community is going to be impressed with initially – unless you work with them initially. Then it becomes that kind of collaboration you want.

That is really the only thing I can think of, as a direction, that seemed not altogether missing here, but I think that people are craving more of it even in their own work.

Lucja Biel: Drawing on what Larry has just said: We have to think of ways how to increase the uptake of corpora by the legal community, because right now we know how to use it for research. We have some applications for teaching students for example, we have corpora for training translators. However, there is still a problem with the uptake of corpora among the professionals, especially in the legal field. I think it would be in-

teresting to invite more lawyers to collaborations, to see what they need, and what they expect, so that our research can be more meaningful to them.

Andreas Abegg: I very much support this. I found it fascinating to work with a linguist, because he knew the tools or methods and I thought about the relevant questions that could be asked. Such a collaboration can be a very fruitful, very creative work, in course of which many new research questions may be discovered.

Hanjo Hamann: To relate to that, yesterday's second presentation, by Stephen, nicely told the story of how he basically brought corpus linguistics into the courts, which is a good illustration of that: 'I told a judge that this [corpus linguistics] is there, I told him how to do it.' – and this basically set off an entire avalanche of work in legal practice. I take it there are at least two people here who will attend the conference at BYU in April on "Corpus Linguistics and the Law", and as far as I know, the roster of participants contains a lot of lawyers who probably haven't done a lot of corpus linguistics. My hope is that this will influence legal scholars and judges in the US. For us Europeans it is easy to take it once Americans have taken it up, because then our research institutions will gratefully fund things. All the originality that comes from Europe aside, there is still a heavy dependency on role models, I guess. Americans are often role models in what they do, so we have always looked to them after the Second World War, because their research is often ahead by ten or twenty years. So I think one path this will go through is through American lawyers taking it up in their court decisions, and German and European lawyers see that and transfer it to their domain. In Europe we have to try to inspire judges and lawyers, which is not as easy, because they are not as open to social science matters in general and linguistics in particular.

Dieter Stein: You know, I have been trying to persuade our Düsseldorf lawyers to come and let me do service for them. Linguists are in a position where they are often talking to lawyers: "Why do you not come and love us?" The thing is, there is a wall between lawyers and linguists, and the name of this wall is ideology. It is an ideology of language. This is what I find very hard: To persuade lawyers to not pursue their ideology of what language is, what words are, what meanings are, and so on. I think there are two ways of handling this. One way is to try and educate lawyers. That is totally futile. Do not even try. The other way for us is to condescend and say: "Okay, we try to speak your language." That borders on prostitution in a way, does it not? [Addressing a lawyer in the audience:] What is your impression, as a lawyer: Am I misrepresenting you?

Ralph Christensen (Mannheim): No, no, you don't. You have to start the conversation. It started in Germany. Brothers Grimm were both lawyers and linguists. And now the lawyers have the power and the linguists are the ones who have the knowledge…

Dieter Stein: They have got the guns!

Ralph Christensen (Mannheim): … and we have to get these back together.

Friedemann Vogel: But I think this is a second problem. It is not only the differences of language ideology, but also law is connected to power. Linguists and social scientists explore what lawyers do, and claim they could make it better, maybe, and this is not only a question of methodology. The question is: Who can speak with whom about power? And law is power, it is the fundamental structure of sharing power and controlling power. So I would be interested what you think: If lawyers came to me and told me 'Nice work, but show us what you have. Here you have to be more normative. I will show you a better method.' and so on – I would not be all too happy about that.

Andreas Abegg: I am not very worried here, because there are so many different levels we could collaborate and benefit each other. If a linguist would come and describe with my corpus how the language developed and how the court used words, I would be fascinated. That would be a contribution to legal theory. It would probably not be taken up by the Federal Court, nor by an article, arguing how you should construe some kind of statute. But so what? It would still be very valuable.

But then again we have those topics where really both disciplines align, as Stephen has told us with the example of the grammatical interpretation. There we are very close. I could think about the use of words and patterns for example. That is a very direct use. We have both our competences aligned directly. And I think we have to try to be creative and then find other ways to collaborate, other topics. It is a fascinating time. I feel that everything is possible at the moment.

Friedemann Vogel: This is the question. Is really everything possible? Or is it – in a critical view – only a game, where we can play, play with legal texts, but with no impact on the practice, where power is made and reproduced? Look to Europe at the moment. We are in a really difficult situation, and nobody knows what the results will be in the next years. What could we do? Could we contribute in this situation with our perspective?

Dieter Stein: We are convinced we could and we should. In fact, we must. But they have to let us, you see.

Larry Solan: I have to say that, at least with respect to translation theory, the EC spends something between 500 and 600 million euros a year on translation. You talk to the translators and they feel like they are sitting in a room with no windows. The translators all feel oppressed. Everybody wants more efficient translation and everybody wants translation where you do not have too many legal problems. The truth is, at least with the regulations and directives, you really do not have that many legal problems. You do not have that many cases coming up where that is a big issue. You probably have them coming up in international courts and nobody notices it. That probably happens. You do not find them in the court of justice of the EU. You find eight cases a year, or something like that. That is not many for a big society like this.

So when you find both, the resources from a linguistically trained group and the corpus perspective, infiltrates only to some extent, but I am talking more generally. Society is feeling the need to have more sophistication with respect to language analysis. I think these people have a fair amount of influence on translation procedures, and there is serious debate about it and there is much less of a gap between conferences, legal translation within Europe and the people who consume it, namely the commission. There are always people from the European commission at these conferences, they are often the keynote speakers. There is a real collaboration in an area like that.

Then you get the statutory interpretation or the interpretation of contracts. The Americans have these weirdest traditions with their dictionaries. I remember once I was consulted on a contract case. The lawyer said: "So I have this linguist, he is a law professor, he gets…" – "I do not need your Henry Higgins telling me how to speak English!" So that is the power relationship that you have, and that happens. You really do not want linguists in court every time anybody is having a dispute about what a law means. You need to hire a bunch of linguists, and you probably do not. So you need to learn the right way to take care of exactly these unusual cases, but they are not totally rare, they are just once in a while.

To me, the challenge here is what you were talking about, Stephen. It is about replacing looking at six different dictionaries, which are just the luck of the draw given that these are all borderline cases of concept formation – who knows which one is going to hit the jackpot, for this side or that side –, and replacing it with the legal system taking lexicography seriously through the data that you have, which everybody in this room knows how to use well. It is about substituting good analysis of word usage for a snapshot that a lexicographer wrote when they are given an average of three lines per word and they plagiarize anyway. If that substitution succeeded, that would be really helpful.

It looks like with these conferences, like the one that BYU is running, there is some chance at least over there of it happening and conceivably, then coming over here and talking to lawyer groups. It can spread in a way that Hanjo suggested. At least that point can. The translation points can. I think assisting in legislation already is happening here. So it happens opportunistically. We need to identify the problems within the legal system in a way the legal system will find it welcoming. That seems to be going on to a greater or lesser extent, depending upon the project.

Dieter Stein: *I think it is interesting that this should develop into a discourse on power, ultimately.*

Friedemann Vogel: This is the point. Power is the point that is important for me. I wonder if it would be more than a symbol to ask or to create an internationally united European corpus of legal language. Do we need such? And how could we proceed?

Dieter Stein: *This would be a top-down interest, in fact.*

Friedemann Vogel: And it is obviously relevant. What do you think?

Larry Solan: That is a perfect example of a project that really needs collaboration from the beginning. Without it, it is a big risk. With it, it is something that is still a risk, but not as much anymore.

Dieter Stein: So this will be one of the take-home bottom lines from this conference that we could subscribe to. And the other is: My impression is that all the lawyers feel they are being, in a way, deconstructed if linguists talk to them. Don't you agree? […]

Stefan Höfler: If I may contradict you a little bit, I am not quite convinced. I am not sure if I buy into this argument about power. I am not sure if it is really a matter of power. I'd rather say that it is a matter of communication. In my experience, what is important is that we, as linguists, do not go to lawyers and tell them where their problem is. First, we have to listen to them and try to figure out where they think their problem is. And then we will see whether we can support them. That is not a power struggle, but a struggle for communication and for trying to understand each other's worlds. I personally think this is a much more fruitful way of looking at the situation.

Dieter Stein: Can I just support you? The godfather of one of my children is actually a lawyer. I was trying to discuss these issues with him. He replied: "*Was wollt ihr denn? Es läuft doch!* – What the hell do you want? It is all alright. We do not need you, really. What is wrong with us?"

Stefan Höfler: Let me look at the situation from my perspective of the problem, legislative drafting: If I go to a lawyer and if I tell them that they should really write clearer laws, because everybody should understand them, then obviously the lawyer would say: "Bugger of!" So that argument does not work.

Dieter Stein: That is a democratic Swiss concern: "*Populärdemokratie*" [direct democracy]. We do not care in Germany.

Stefan Höfler: But if I go to the same lawyer and I explain to him, how his own work will become easier, because a statute is written in a clearer way, then he will be much more open to my suggestions and that is the way I think we should go forward.

Dieter Stein: I think we all think that way.

Victoria Guillén-Nieto (Alicante): I completely agree with you. I have been attending this conference and I think it is fascinating. As for the methods that are applied, and the way the corpora are build, it is really wonderful. But I can see some weaknesses.

The first weakness is that even if we joined together in the creation of an international legal corpus – which I am very much in favour of – what is the purpose? What is the purpose of gathering a corpus, what do we need this corpus for? Since we moved into this society of information and knowledge, the way we set our hypotheses is not

just academic. It has to be socially and professionally relevant. 'Begin at the beginning', the King said bravely, 'and when you come to the end, then stop.' This is Lewis Carroll in Alice in Wonderland. But I do think the "first thing" is to organise a group together with professionals and listen to them. And then, once we listen to them, we can really brainstorm and gather lots of ideas. And then we can focus on the sort of corpora we need to build and the sort of hypotheses we need to set up to make sure that apart from being academic, they are professionally and socially relevant. Otherwise the findings, whatever we do and however great and wonderful it is, will not be transferred to the society of information and knowledge.

If we establish this relevance, we can also get funds. I can tell you about an experience I had years ago, which had nothing to do with the legal language, but it had to do with intercultural pragmatics. Begin at the beginning, we organised a group, together with people who were in international business, we found out the red lights, we did research but at the same time, the findings of this research were transferred into the creation of a graphic adventure. The graphic adventure thing attracted 60,000 Euro, which is something that no one at the faculty of humanities at the university Alicante had dreamt of. It was relevant. It was academic, but it was socially and professionally relevant and we think: 'Begin at the beginning' is to group together with professionals. This will really help us to find the purpose, because we already do very well in the methods and in the building of corpora.

Dieter Stein: *I think here is another take-home message.*

Victoria Guillén-Nieto (Alicante): And I am very much in favour of organising this international team and constructing an international corpus…

Łucja Biel: … that will let us look for patterns above specific national languages. […]

Dieter Stein: It would be something like the successor to Eurotyp. Remember that, Eurotyp?

Hanjo Hamann: I agree with that, but I also want to put it in perspective, because I think the way corpus linguistics has proceeded in many cases is top-down, in some cases with questions. Then you assemble your corpus and then you throw the corpus away and it is forgotten. What I think is missing is an infrastructure and you cannot define exactly the task of an infrastructure because there are so many ways to use it. Whatever I think of the EUR-Lex collection of documents at the European level, this started out as an infrastructure to make legislation public and the EU transparent. And I think meanwhile it has found so many applications in corpus linguistics and law and other areas, which were never intended. It was always 'Oh, there is this material, what can we do with it?' – and suddenly you find a wealth of research questions that you can answer with that. And in that sense I think, even building a corpus without some specifications of questions, can be useful as an infrastructure.

What I find most challenging in our projects is that all databases that we as lawyers have are good as long as you 'do it the way they always did'. That is: I look for a single document, which I want to read. How can I get to it most quickly? And they [our current databases] are good for that. They are appropriate. But they are never set up in a way to look at a number of documents in a general frame 'from above'. Take as an example the things that I showed in our presentation about the quality of the juris data: I said there are something like eleven court decisions that have wrong page numbers. That is in a universe of 9,000 court decisions. Nobody would care if we did not do it from a bird's-eye perspective. But being able to change the focus from 'databases are just for retrieving documents that you can then read', to a perspective like 'databases are collections that you can look at on a microscopic level' is something that is missing entirely from current databases. That is a sort of infrastructure that we will need. For example, if EUR-Lex came with a KWIC, a keyword-in-context display with collocates and everything: That would help in so many ways, even if we do not have the research question yet to address with this. But just extending the infrastructure so we can do these things at all, I think, would be helpful.

Dieter Stein: *Thank you very much. I think this is a wonderful concluding statement with practical advice. Our time is nearly up. This is the time, as we were talking about power, to use my own personal power to thank you for putting up this wonderful conference, and thank you for your contributions.*

Have a safe journey home!