RiCL Research in Corpus Linguistics

A lexico-grammatical comparison of statutory law and popular written language

Margaret Wood
Northern Arizona University / United States

Abstract – While the plain language movement has shed light on the lack of readability of statutory texts for the lay person, there has been a lack of empirical methodology employed to determine the ways in which statutory language differs lexico-grammatically from forms of popular language that are familiar to the lay person. With this in mind, the present study conducts a comparative analysis of statutory language and other forms of popular written language (i.e., a corpus of news reports, sports reports, encyclopedia articles, and historical articles) with two goals: 1) to provide a detailed lexico-grammatical description of statutory law independent from other forms of legal writing, and 2) to identify pervasive lexico-grammatical features of statutory language that the lay person has relatively less exposure to in comparison to other written registers. Following a bottom-up selection of lexico-grammatical features for analysis, a key feature analysis is used to identify linguistic features that are more pervasive in statutory law relative to other forms of popular written language as measured through Cohen's *d* effect sizes. Results reveal the pervasive use of the passive voice, prepositions, a variety of coordinating conjunctions, the pied-piping *wh*-relative clause construction, and non-finite *-ing* and *-ed* clause constructions in statutory language. These results complement previous research regarding the features that are characteristic of statutory language and help to identify features that potentially contribute to the lack of readability of statutory law.

Keywords – statutory law; register variation; readability; popular language; key feature analysis

1. INTRODUCTION

For years, people have bemoaned the lack of readability of written legal documents, in particular for those outside of the profession or without detailed knowledge of the law. The plain language movement, which has its roots in the 1970s, calls for legal language that is accessible and readable for the lay person. With this has come numerous attempts to describe the language of written legal documents and identify the features that are detrimental to the readability of the texts.

While these linguistic descriptions have concerned a variety of written legal texts (all of which pose readability challenges for the lay person), there are comparatively fewer empirical descriptions of statutory language independent from other forms of legal writing (e.g., contracts, agreements, treaties). The current lack of independent

focus on statutory law is problematic, as the domain carries an extraordinary amount of power over the lay person; explicitly creating, modifying, and terminating legal rights and obligations of everyday individuals (Tiersma 1999: 1). Because a long history of register variation studies tells us that linguistic characteristics of a text will differ in relation to the situational context in which they occur (Biber and Conrad 2009), it is important for the discussion of readability of legal texts to identify the lexico-grammatical features that are pervasive in the register of statutory law as an independent form of written legal language.[1]

Claims are frequently made about the pervasiveness of certain features of legal writing based on simple frequency counts within a register or across multiple combined registers. While this may tell us which features are more common in the register relative to other features in that same register, if we wish to identify features that are uniquely characteristic of statutory law and aim to make claims about pervasiveness, the register must be described in relation to a different text variety or domain. Using non-legal language as domain for comparison stands to contribute to the discussion of readability as it allows for the identification of pervasive lexico-grammatical features in statutory law that the lay person has relatively less exposure to on an everyday basis. The value of this lies in the assumption that a lack of exposure to the characteristic linguistic structures of a specific text variety has the potential to impede one's understanding of it. We see evidence of this in the fact that the typical 'audience' of statutes, or those who interact with them on a daily basis (i.e., lawyers and judges), seem to be able to make sense of the texts more readily than the lay person.

With this in mind, the present study aims to provide a linguistic description of codified state statutory law in relation to other forms of non-legal, popular written language, with two goals: 1) to provide a detailed lexico-grammatical description of the features that are characteristic of state statutory law (as a text variety that holds great power over the lay person), and 2) to identify pervasive lexico-grammatical features of statutory law that the lay person has markedly less exposure to in comparison to other forms of popular written language. The present study proposes that the bottom-up (rather than top-down) identification of features that are pervasive in statutory law and relatively less common in other forms of popular written language will allow for the

[1] The present study uses the term 'register' to refer to culturally-recognized text varieties (Biber and Conrad 2009: 6).

removal of personal intuition concerning which features are relevant in the conversation of readability. Through a linguistic comparison of a corpus of state statutory law (i.e., bills proposed by an elected member of a state house or senate, drafted by a draftsperson, passed through various committees, and signed into law by the governor of the state), and a corpus of popular written language comprising online news reports, sports reports, encyclopedia articles (i.e., *Wikipedia*), and historical articles, the present study aims to contribute to the discussion concerning the language that poses a threat to the readability of statutes for the lay person.

2. LITERATURE REVIEW

2.1. Linguistic descriptions of written legal language

Previous linguistic descriptions of written legal language have concerned a variety of registers including decisions, directives, regulations, law journals, commercial law documents, case law, contracts, law reports, and legislation. Linguistic studies of these registers have most commonly focused on lexis: in particular, lexical bundles and keyword analysis (Caliendo *et al*. 2005; Trebits 2009; Jablonkai 2010; Breeze 2013; Biel 2017; Alasmary 2019; Serachini 2020), phraseology (Biel 2009, 2014; Pontrandolfo 2015) and on single features such as modals (Foley 2002; Andersson 2007; Gibova 2011) and personal pronouns (Rodríguez-Puente 2019).

Only a select number of studies that have described forms of written legal language in terms of their lexico-grammatical characteristics, though these have largely been undertaken without the use of a reference register (notable exceptions include Goźdź-Roszkowski 2011, and Biber and Gray 2019). Studies that have focused on the lexico-grammatical characterization of legislative writing have described it as both structurally compressed and structurally elaborated. The frequent use of nominalization (nouns that have been morphologically derived from verbs or adjectives), which are features often associated with structurally compressed written language (Biber 1988; Biber and Gray 2016), are considered highly characteristic of legislation (Goźdź-Roszkowski 2011; Sun and Cheng 2017). Williams (2013: 354) similarly characterized legislative language as structurally compressed, noting in particular its reliance on nouns, including the frequent use of nominalization and high density of noun phrases.

Others describe structural elaboration of legislative language through the density of clausal embedding, which is frequently considered one of the most detrimental features to readability (Williams 2007). Charrow and Charrow (1979: 1329) specifically attribute readability issues to central embedding, in which there are two subordinate clauses; one enclosed within the other. Bhatia (1983: 50) also noted that legislation displays a high degree of subordination, citing adverbials and non-finite prepositional constructions as particularly common. Embedded clauses have been referred to as 'qualification inserts', which are used to flesh out main ideas of a clause and directly contribute to the syntactic complexity of legislative language (Bhatia 1993). Goźdź-Roszkowski (2011: 136) found that legislative language made particularly frequent use of different types of post-nominal clauses, including *wh*-relative clauses, *that* relative clauses and the pied-piping construction. Tiersma (1999: 62) also noted that legislation frequently makes use of coordinating conjunctions *and* and *or* to combine multiple clauses, contributing to the 'wordy' nature of the texts, and states that "the possibilities of creating tremendously long phrases and sentences by use of conjunctions like *and* and *or* are virtually limitless."

Use of the passive voice is also considered highly characteristic of legislative writing. According to Williams (2004: 231), approximately one quarter of all verbal constructions in prescriptive legal English are in the passive voice. Bulatović (2013: 103) found that of the verb phrases counted in a corpus of acts, around 65 per cent were in the active voice and 35 per cent were in the passive voice. Of those passives, around 24 per cent served as post-nominal modifiers in the form of past participles (Bulatović 2013: 104).

However, as previously noted, a majority of the studies above describe legislative writing without comparison to other registers. It is difficult to know, for example, how notable it is to have a text with 35 per cent of its verbal constructions in the passive voice, if there is nothing to compare this percentage too. For this reason, the present study aims to test these claims about pervasiveness through empirical, comparative means.

2.2. Linguistic descriptions of popular written varieties

The present study focuses on written language that is 'popular', that is, on language that is written specifically for the lay person as its audience and is easily accessible to them. For this reason, the study investigates the online popular written registers of news reports, sports reports, encyclopedia articles, and historical articles as registers that fit these criteria (see Section 3.1.2).

The most prominent large-scale linguistic description of forms of popular language was undertaken by Biber *et al.* (1999) in the *Longman Grammar of Spoken and Written English*. Using the *Longman Spoken and Written English Corpus* (LSWE), which comprises over 40 million words representing six registers, Biber *et al.* (1999: 5) compiled a "descriptive and explanatory account of English grammar." Four core registers were used in their analysis: conversation (British), fiction (American and British), news (British), and academic prose (American and British). Biber *et al.* (1999: 25) also included two other sets of texts for dialect comparison (American conversation and American news), and two supplementary registers (British non-conversational speech, and British and American general prose). Biber *et al.* were able to investigate structural descriptions of the features and patterns of use, and comment on the pervasiveness of the features in comparison to other registers. They undertook extensive functional interpretation of the quantitative data, in particular in terms of three functional associations: the work that a feature performed in discourse, the processing constraints that it reflected, and the situational and social distinctions that it conventionally indexed (Biber *et al.* 1999: 41). Of particular interest to the present study is the lexico-grammatical comparison of formal academic prose to other non-academic registers, as academic prose generally shares much in common with previous descriptions of legislative writing, namely, the tendency towards dense, informational, compressed language.

The popular written register of news has frequently been the subject of investigation, largely studied through discourse analysis (e.g., Davies 2012; Fowler 2013; Bednarek and Caple 2014; Scollon 2014; Xie 2018). These studies have often focused on highly specific contexts; for example, political posts in the Jakarta post newspaper (Yana 2015) and socio-political influences on lexico-grammatical features in Ecuadorian Spanish news (Tapia and Biber 2014). However, select others have had a broader focus. In a multi-dimensional analysis of registers on the searchable web, Biber

and Egbert (2016: 109) found that news reports were characterized by a set of co-occurring features frequently associated with written informational language; largely, a variety of nominal modifiers. Biber and Egbert (2016) also found, however, that news was characterized by the co-occurrence of features such as complement clauses and *that* deletion, which are often associated with oral language varieties. Through a later key feature analysis in *Register Variation on the Web*, Biber and Egbert (2018) found that when set aside a reference corpus of other web registers, news reports displayed a relatively higher use of communication verbs, proper nouns, common nouns, perfect aspect, pre-modifying nouns, and prepositions.

In both studies, Biber and Egbert (2016, 2018) provided linguistic descriptions of a variety of other web registers, including encyclopedia articles, historical articles, and sports reports (registers of analysis in the present study). Biber and Egbert (2016) found that encyclopedia articles were characterized by the co-occurrence of features associated with literate-informational language (prepositional phrases, passive non-finite relative clauses, relative clauses). In the later key feature analysis, Biber and Egbert (2018:162) found that passives, prepositions, longer word length, and nominalizations were key in the register. They found that historical articles had similar key features, though with the notable added use of the past tense, which was the most key feature in the register with a large effect size of $d > 1.0$ (Biber and Egbert 2018: 95). On the other hand, sports reports made pervasive use of features associated with narrative and oral varieties when set aside a reference corpus of the web registers, including proper nouns, third-person pronouns, activity verbs, past tense, perfect aspect, contractions, and adverbs of place (Biber and Egbert 2018: 90). Notably, while proper nouns were also key for news reports, the effect size was more than two times larger in sport reports (Biber and Egbert 2018: 91).

Largely influenced by the work of Biber and Egbert (2016, 2018), the present study combines several of these web registers in order to build a reference corpus representing popular written language as a whole. This has been done in order to increase coverage of the various types of online language that individuals have frequent exposure to.

2.3. Comparisons of legislative language and non-legal language

While the literature discussed in Section 2.1 has constituted a great contribution to our knowledge of legislative language, the prevailing gaps remain: 1) a focus on legislation independent from other written legal language, and 2) a lexico-grammatical description of statutory law in reference to other types of language. To the best of the researcher's knowledge, only two studies have made empirical lexico-grammatical comparisons of legislative writing and non-legal registers. Goźdź-Roszkowski's (2011) register variation study of legal language was undertaken with the goal of comparing a variety of legal registers to one another, including academic journals, briefs, contracts, legislation, opinions, professional articles, and textbooks. While the primary goal of Goźdź-Roszkowski's study was to examine lexico-grammatical variation between legal registers, he briefly compares the seven legal registers to select forms of non-legal language (i.e., fiction, textbooks, conversion, research articles, academic prose) through an additive multi-dimensional analysis on Biber's (1988) dimensions. In doing so, Goźdź-Roszkowski characterized legislation as comparatively informational, non-narrative, explicit (as opposed to situation-dependent), and lacking overt persuasion. Also of importance for the present study is the considerable amount of variation that Goźdź-Roszkowski found *between* legal registers, lending further support for the argument that for a clear and accurate description of a particular type of legal language, one must study it as a unique, independent register.

The other study that has undertaken a comparative lexico-grammatical analysis of legal and non-legal language was conducted by Özyildirim (2011), who investigated Turkish legislation in relation to other forms of non-legal language, including Turkish scientific research articles, newspaper articles, television commercials, men's/women's magazines, and stand-up comedy shows. Özyildirim (2011:78) made use of an additive multi-dimensional analysis on Biber's (1988) dimensions as Goźdź-Roszkowski did, but focused only on the narrative vs. non-narrative dimension, similarly characterizing legislation as highly non-narrative.

In some ways, this analysis follows in the footsteps of Goźdź-Roszkowski (2011) and Özyildirim (2011), though the present study differs both in methodology and research aims. First, both Goźdź-Roszkowski and Özyildirim made use of a multi-dimensional analysis for their register comparisons, which is used to characterize a number of individual registers in terms of the co-occurrence patterns of lexico-

grammatical features. In contrast, the analysis here focuses on identifying features that are markedly pervasive in one register relative to a combined reference corpus of other registers and does not concern feature co-occurrence. Finally, the selection of non-legal registers is targeted specifically for the purposes of investigating readability. While both Goźdź-Roszkowski and Özyildirim used a mixture of spoken and written registers as well as academic registers (i.e., textbooks and research articles, which are not considered 'popular' in the present study due to the restricted audience), this study makes use of a much more narrowly defined group of texts, specifically representing language that is both accessible and familiar to a lay audience.

3. Methodology

3.1. Corpora

The present study makes use of two corpora for analysis: a corpus of state statutory law, and a corpus representing other forms of popular written language. The following sections will describe the motivation for selection of the text varieties in the two corpora and the compilation processes.

3.1.1. Corpus of state statutory law

The corpus of state statutory law used for the present study was sampled from the larger *Corpus of United States State Statutes* (CorUSSS) (Egbert and Wood under review), which comprises the state codes for each of the 50 states in the United States. CorUSSS was compiled using a *Python* script to web-scrape texts located on https://www.justia.com. Statutes were initially scraped and aggregated at the top level by title, each of which contains a set of statutes representing specific topical content (e.g., Agriculture, Criminal Code, Businesses, Corporations). Text files were cleaned through a second *Python* script that removed all boiler-plate text and inserted brackets into the text files to denote meta-data, including the name of statute, year, and universal citation. A secondary cleaning process was undertaken through the regular expression program *Sublime Text* in order to remove extraneous boiler-plate text leftover following the initial cleaning process.[2]

[2] https://www.sublimetext.com/

To compile the corpus of statutory law used in the present study, a sample of eight state codes was selected from the 50 states. This smaller selection was made for logistical reasons, namely, any linguistic analysis on a corpus of such size (the totality of CorUSSS consists of over 420 million words and almost 8 million texts) would be challenging to conduct with existing corpus analysis tools. Additionally, because the compilation process of a corpus this large was fairly time-consuming, only a limited number of the state codes were available for use (web-scraped, cleaned, and tagged) at the time the present study was carried out. However, during the design and construction of CorUSSS, exploratory frequency counts of a variety of linguistic features in these states revealed very little variation between the codes from state to state, providing a high level of confidence that even if a complete corpus of all 50 state codes was used, there would not be substantial changes to the results. Still, in the selection process of state codes available at the time of the study, care was taken to select state codes that represented a variety of geographical regions in the United States in order to control for representativeness of the country as closely as possible. The final selection of states resulted in a corpus of state statutory law comprising 670 texts and 90,388,372 words. The final composition of the corpus is presented below in Table 1.

Codes	**Number of texts**	**Number of words**
Rhode Island	155	6,190,952
West Virginia	133	6,952,846
Kansas	85	5,795,347
Connecticut	72	7,798,889
New Jersey	68	10,855,203
South Dakota	68	4,210,208
South Carolina	63	5,993,304
Alaska	43	873,860
Total	**670**	**90,388,372**

Table 1: The statutory law texts

The texts were tagged for lexico-grammatical features using the *Biber Tagger*, which identifies a larger set of characteristics than other existing taggers (over 150 features) and is able to identify these features at a more fine-grained level, for example, the identification of the gap position for *wh*-relative clauses (Biber and Egbert 2018: 22). Staples *et al.* (2016) reported that the tagger tagged at 90 per cent accuracy for formal writing.

3.1.2. *Popular Written Language* corpus

The *Popular Written Language* corpus (PWL) used for the present study comprised a selection of web registers. This decision was made based on the criteria that language needed to be written for an audience of the general public, and easily accessible to that population. Because the Internet is highly accessible to the general public in the United States (whether personally or in public establishments) and reaches a wide audience, registers selected to represent popular written language were sampled from the *Corpus of Online Registers of English* (CORE). CORE is a corpus compiled by Biber and Egbert (2016) sampled from the larger *Corpus of Global Web-based English* (GloWbE; Davies 2013). The entirety of CORE holds 48,571 documents and nearly 54 million words (Biber and Egbert 2016: 14). Using CORE was also beneficial as Biber and Egbert (2016) had previously removed any texts from the sample that had fewer than 75 words, which is undesirable for studies of lexico-grammatical characteristics (Biber and Egbert 2018: 13).

Popular written registers were selected from CORE with the aim of keeping the PWL corpus as cohesive as possible in terms of situational characteristics. To be included in the PWL corpus, registers needed to be written by an author that has formal expertise or insider knowledge of the topic about which they are writing (i.e., news, sports, history, etc.). Registers were not selected for the corpus if they varied in mode (i.e., spoken language), were not written for a lay audience (academic research articles), did not represent real-world topical content, or were highly stylistically varied (i.e., fiction). To be included in the corpus for the present study, registers also needed to be originally in the written mode and be non-interactive (categorized as such by Biber and Egbert 2018).

Appendix 1 provides an overview of situational characteristics for all five registers used in the present study, demonstrating the relative similarity in most of their characteristics. The situational difference between these registers lies predominantly in topic, with a small range of variation in communicative purpose.

Extensive consideration was given to blogs, which were selected for the corpus in the early stages of the project due to the popularity of the text type. However, Biber and Egbert (2018) identified this text type as one that does not seem to clearly fit a register, as topic and blog type are highly variable. In the end, blogs were excluded from consideration with the exception of two types: sports blogs and news blogs. This

decision was made for two reasons. First, these two types of blogs are infrequently written by the lay person, but rather individuals with relatively specialized knowledge of the topic. Along with this, they infrequently concern personal experience, instead reporting on outside stories or occurrences. This is in contrast to other blog types identified by Biber and Egbert (2018), such as personal narrative blogs, travel blogs, and opinion blogs, all of which were excluded from the PWL corpus. Second, Biber and Egbert (2018: 42) chose to incorporate news blogs and sports blogs to their respective registers based on the finding that often these blogs were "virtually indistinguishable from published reports." The final composition of the PWL corpus is presented in Table 2.

Codes	**Number of texts**	**Number of words**
News	600	498,780
Sports reports	600	472,795
Encyclopedia articles	430	1,291,380
Historical articles	206	413,537
Total	**1,871**	**2,756,389**

Table 2: The *Popular Written Language* corpus

3.2. Linguistic analysis

3.2.1. Key feature analysis

To identify pervasive lexico-grammatical features in statutory law, the present study makes use of a key feature analysis. Key feature analysis makes use of a reference corpus in order to identify features that are markedly more frequent in a target corpus, which are considered 'key'.

Key feature analysis makes use of the mean rate of occurrence and standard deviations of linguistic features to calculate Cohen's *d* effect sizes (Cohen 1977). Large positive Cohen's *d* values indicate that the feature is markedly more frequent in the target corpus than in the reference corpus, while large negative Cohen's *d* values indicate that the feature is markedly less frequent. In accordance with Cohen (1977), *d* values will be interpreted as small (> +/- 0.20), medium (> +/- 0.50) and large (> +/- 0.80).

In the present study, features with large positive effect sizes in the corpus of statutory law are considered pervasive linguistic features of statutory language that the

lay population is expected to have less exposure to on a daily basis. Cohen's *d* values approaching zero are an indication of a similar frequency of use in the two corpora.

3.2.2. Feature selection

The lexico-grammatical features selected for analysis were generated through bottom-up means in order to remove the influence of personal intuition concerning the pervasiveness of certain features. The general tag count generated from the *Biber Tagge*r was used, which provides normed frequency counts per 1,000 words for over 150 lexico-grammatical features. A normed frequency count for one additional lexico-grammatical feature —the non-finite post-nominal *-ing* clause— was manually added.

Once normed frequency counts for all features were obtained, a dispersion threshold of 90 per cent was established, meaning that the feature had to appear in at least 90 per cent of the texts in either of the two corpora in order to be retained for analysis. This narrowed the list of features for analysis to 81 features. An additional 19 features were then eliminated from the analysis due to overlap. This included the removal of several 'all' features (e.g., 'all adjectives'), in favor of more specific types of that feature (e.g., 'predicative adjectives' and 'attributive adjectives'). Specific semantic domains were later removed if they included similar lexical items; for example, private verbs and mental verbs, which include words such as *think* and *believe.* In such cases, the semantic domain with the larger effect size (positive or negative) was retained for analysis. This resulted in a final list of 62 linguistic features, which are presented in Table 3.

Verbs	**Dependent clauses**
Present tense	Non-finite *-ing* clauses
Past tense	Non-finite *-ed* clauses
Perfect aspect	*To* complement clause controlled by verbs of modality, causation
Progressive aspect	*To* complement clauses controlled by stance nouns
Passive + *by*	*That* complement clause controlled by verbs
Passive post-nominal modifier	*That* relative clauses
Short passives	*Wh* relative clause, object position
Infinitive	*Wh* relative clause, subject position
Split auxiliary	*Wh* relative clause, prepositional fronting (pied-piping)
Be as main verb	
Have as main verb	**Other**
Modals of prediction	Stranded preposition
Modals of possibility	Prepositions
Mental verbs	Clausal coordinating conjunction
Communication verbs	Phrasal coordinating conjunction
Activity verbs	Subordinating conjunction – conditional
Suasive verbs	Subordinating adverbial - other
Aspectual verbs	Attributive adjectives
Verbs of likelihood	Predicative adjectives
Verbs of existence	Linking adverbials
Verbs of causation	1st person pronouns
Verbs of occurrence	3rd person pronouns
	Indefinite pronouns
Nouns	Pronoun *it*
Process nouns	Indefinite articles
Abstract nouns	Definite articles
Human nouns	Contractions
Place nouns	Topical adjectives
Technical nouns	Adverb of time
Cognitive nouns	Adverb of place
Quantity nouns	Downtoner
Group noun	Type/token ratio
Proper nouns	
Pre-modifying nouns	

Table 3: Features for keyword analysis

4. RESULTS

Results from the key feature analysis are presented in Figure 1 and Tables 4 and 5 below. Figure 1 shows an oral/literate divide between the PWL corpus and the corpus of statutory law which, as suggested by Biber (2014), is a universal dimension in multi-dimensional analysis studies. More specifically, popular written language displays the lexico-grammatical characteristics that are highly typical of narrative language (i.e., first and third-person pronouns, past tense, perfect aspect, progressive aspect, contractions, verbs), while statutory language can be characterized as highly detail-oriented, dense, and topically narrow (type/token ratio was key with a large effect size in the PWL, indicating high lexical diversity relative to statutory language).

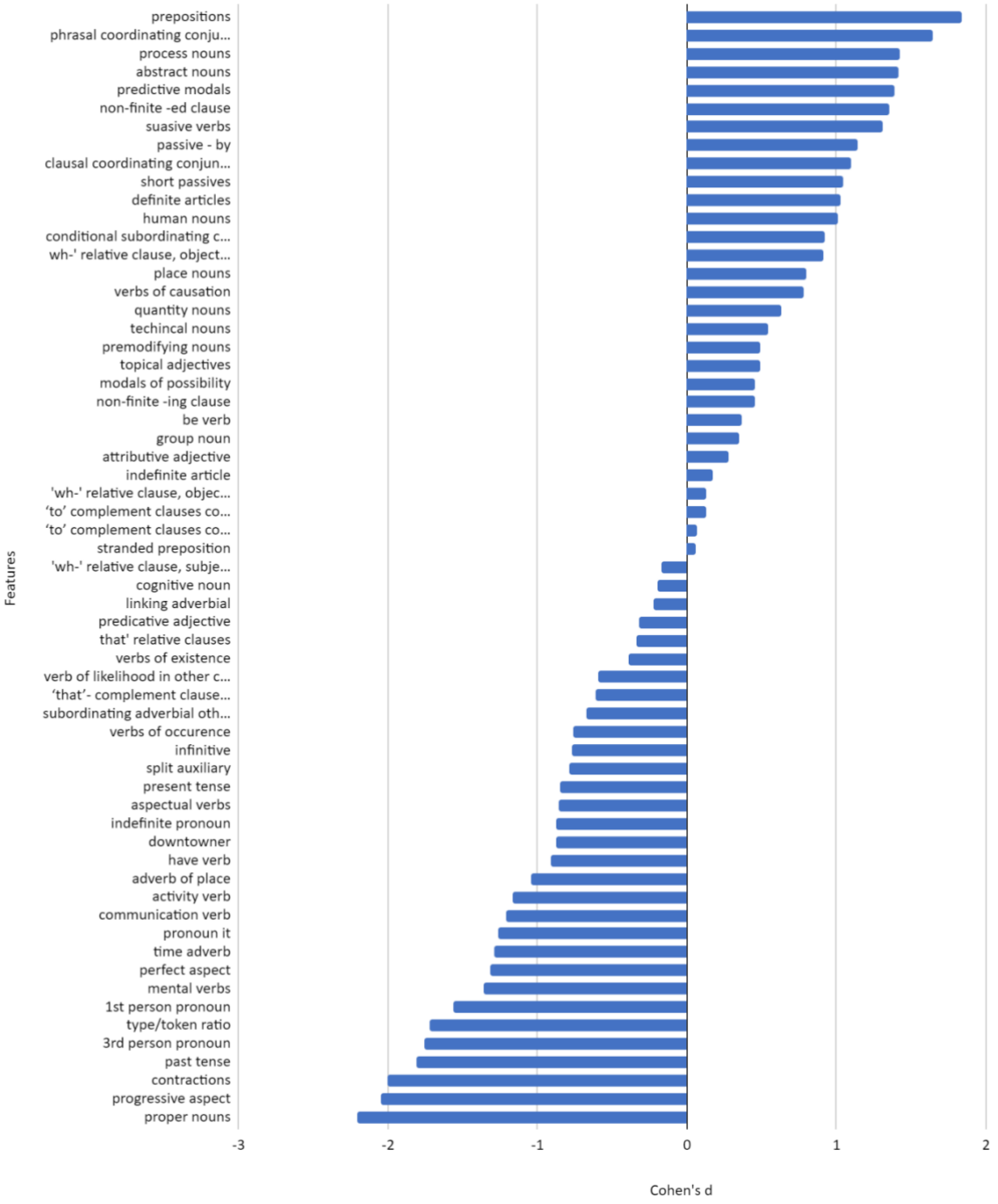

Figure 1: Key feature analysis results

Key features of statutory law indicate the pervasive use of both phrasal language (contributing to the dense, literate nature of statutes) and clausal language (contributing to the long-winded, detail-oriented nature of statutes). The corpus of statutory law has 15 features with large effect sizes over d=0.90. Of these, six features are typically associated with literate language (Table 4). Two passive constructions are key in statutory law with large effect sizes (*by* passives, d=1.14; short passives, d=1.04), as are

prepositions (with the highest keyness score in the corpus of *d*=1.84). Statutory law also demonstrated frequent use of phrasal coordinating conjunctions, which had the second largest effect size in the corpus (*d*=1.64). Nouns of several different semantic domains were also key in statutory law, including process nouns, abstract nouns, human nouns, quantity nouns, place nouns, and technical nouns (with medium to large effect sizes). Key clausal language in the corpus included the use of clausal coordinating conjunctions (*d*=1.10), conditional subordinating conjunctions (*d*=0.92), non-finite *-ed* clauses (passive post-nominal modifier) (*d*=1.35), *wh*-relative clauses with the pronoun in the object position and prepositional fronting (*d*=0.91), and non-finite *-ing* clauses (*d*=0.45). This mixture of both phrasal and clausal features is consistent with past research of legal language previously discussed in Section 2.1. The notable keyness of conjunctions —phrasal, clausal, and subordinating— has been tied to the characteristically long, drawn out sentences found in statutory language (Tiersma 1999).

Effect size	***d***	**Feature**
Large	1.84	prepositions
	1.64	phrasal coordinating conjunction
	1.42	process nouns
	1.41	abstract nouns
	1.39	predictive modals
	1.35	non-finite *-ed* clause
	1.31	suasive verbs
	1.14	*by* passive
	1.10	clausal coordinating conjunction
	1.04	short passives
	1.03	definite articles
	1.01	human nouns
	0.92	conditional subordinating conjunction
	0.91	*wh*- relative clause, object position with prepositional fronting (‘pied piping’)
Medium	0.80	place nouns
	0.78	verbs of causation
	0.63	quantity nouns
	0.54	technical nouns
Small	0.49	premodifying nouns
	0.49	topical adjectives
	0.45	modals of possibility
	0.45	non-finite *-ing* clause
	0.37	*be* verb
	0.35	group noun
	0.28	attributive adjective
	0.17	indefinite article
	0.13	*wh*- relative clause, object position
	0.13	*to* complement clauses controlled by verbs of modality, causation, and effort
	0.07	*to* complement clauses controlled by stance noun
	0.06	stranded preposition

Table 4: Key features for statuary law

The negative effect sizes indicate markedly less frequent use of verb-associated features in the statutory law corpus in comparison to the PWL corpus. Features with medium to large negative effect sizes include: progressive aspect (d = -2.05), past tense (d = -1.81), perfect aspect (d = -1.32), time and place adverbs (d = -1.29; d = -1.04), present tense (d = -0.85), split auxiliaries (d = -0.79), the infinitive (d = -0.77), and a variety of semantic domains of verbs (mental verbs, communication verbs, activity verbs, aspectual verbs, verbs of likelihood, verbs of existence). Typically, narrative features with large effect sizes also included a variety of pronouns (first person, third person, pronoun *it*) and contractions. Proper nouns had the largest effect size in the PWL corpus, which is unsurprising based on the selection of registers in the present study, which can concern an unlimited number of different people and places.[3] Various clausal features also had a large effect size in the PWL corpus, including *that* complement clause controlled by verbs, *that* relative clauses, and the *wh-* relative clauses with the pronoun in the subject position (though the latter three features had small effect sizes).

Notably, features that appeared in less than 50 per cent of the texts in the PWL corpus (and over 90 % in the statutory law corpus) and were key in statutory law with medium to large effect sizes included *wh-* relative clauses with the pronoun in the object position, *wh-* 'pied-piping' relative clauses, and suasive verbs (e.g., *ask*, *command*, *insist*). The low dispersion of these features across the PWL coupled with the high keyness in statutory law makes these highly important features to consider in the discussion of readability.

[3] Also recall that, in Biber and Egbert's (2018) key feature study, proper nouns had a very high effect size in both news reports and sports reports.

Effect size	*d*	Feature
Large	-2.21	proper nouns
	-2.05	progressive aspect
	-2.00	contractions
	-1.81	past tense
	-1.76	3rd person pronoun
	-1.56	1st person pronoun
	-1.36	mental verbs
	-1.32	perfect aspect
	-1.29	time adverb
	-1.26	pronoun *it*
	-1.21	communication verb
	-1.17	activity verb
	-1.04	adverb of place
	-0.91	*have* verb
	-0.88	indefinite pronoun
	-0.88	downtowner
	-0.86	aspectual verbs
	-0.85	present tense
Medium	-0.79	split auxiliary
	-0.77	infinitive
	-0.76	verbs of occurrence
	-0.67	subordinating adverbial other
	-0.61	*that* complement clause controlled by verb
	-0.59	verb of likelihood in other contexts
Small	-0.39	verbs of existence
	-0.34	*that* relative clauses
	-0.32	predicative adjective
	-0.22	linking adverbial
	-0.20	cognitive noun
	-0.17	*wh-* relative clause, subject position

Table 5: Key features for popular written language

5. DISCUSSION

The literate nature of statutory language is seen in part in the variety of semantic domains of nouns that are key in the statutory law corpus, corroborating the findings by Williams (2013), who noted the nominal nature of legislative texts. While there seems to be a large number of key semantic domains of nouns for a register with content that is far more restricted than that of popular language (which has great freedom in topic), the semantic domains represented are clearly oriented towards topics typically discussed in law. These domains include process nouns (e.g., *system*, *meeting*; *d*=1.42), abstract nouns (e.g., *agreement*; *d*=1.41), human nouns (e.g., *person*, *governor*; *d*=1.01), place nouns (e.g., *town, city*; *d*=0.80), and technical nouns (e.g., *jurisdiction*; *d*=0.54*).* Statutes typically contain descriptions of the people, settings, and contexts in which a law takes effect, meaning that the semantic domains named above complement one

another well. Excerpt 1, below, demonstrates the use of a variety of nouns from different semantic domains (in bold) working together to describe people, setting, and subject matter in a highly specific context. In particular, this excerpt uses a large number of abstract nouns, such as *discretion* and *compliance*.

(1) **Excerpt 1**: The **director** shall have **discretion** to assess an administrative **penalty** of not more than two hundred fifty dollars ($250) per **offense** against any **insurance company** that fails to notify the **director** as required in this section. The **director**, in his or her **discretion**, may bring a **civil action** to collect all assessed civil **penalties**. The workers' **compensation court** shall have **jurisdiction** to enforce **compliance** with any order of the **director** made pursuant to this **section**. (R.I. § 28-36-12).

In contrast, popular written language makes a more frequent use of various verb-associated features, which is characteristic of oral and narrative language. Note that while the topical content of the historical article below concerns matters of law (see excerpt 2), the narrative, story-telling aspect of the text is reflected in the use of past tense, perfect aspect, and proper nouns (underlined), in particular, when compared to excerpt 1. While excerpt 2 narrates a historical event, excerpt 3 appears to narrate an individuals' personal thoughts, making use of both present tense and past tense, and the perfect and progressive aspects. The variety of semantic domains of verbs which are key in written popular language is also notable, including mental verbs (*think*) and activity verbs (*spend*, *move*).

(2) **Excerpt 2**: Historical Article. The hearings **had run** for eleven days. The hearing three years earlier to confirm Fortas as associate justice **had run** for three hours. At the beginning of October, Fortas's nomination **went** to the full Senate for a vote. For four days straight, senators **defended** or **lambasted** Fortas until a cloture petition to end the debate **was introduced**. (https://www.neh.gov/humanities/2009/septemberoctober/feature/supremely-contentious)

(3) **Excerpt 3**: Sports Report. The NHL should **step in** and **cough up** a few $$$. **I think** the NFL **helps** out with new stadiums. Bettman **has spent** millions **to keep** a money losing franchise in PHX. He **could spend** a few more **to keep** a money making one in EDM. Even if the league **approved** relocation for the Oilers, there **would be** 8 teams in line to **move** to Edmonton. This **has** nothing to do with the city and everything to do with Katz not **wanting to spend** any of his $200 Billion. (http://ca.sports.yahoo.com/blogs/nhl-puck-daddy/oilers-talking-relocation-seattle-playing-arena-deal-hardball-012114557--nhl.html)

While the preference for verbs is associated with oral and narrative varieties and the dense use of nouns is associated with statutory language, the key feature analysis

showed exceptions to this based on semantic domain. There are two semantic domains of verbs that are key in statutory language with relatively large effect sizes: suasive verbs (e.g., *ask*, *command*, *insist*; *d*=1.31) and verbs of causation (e.g., *let*, *permit*; *d*=0.78). These two domains of verbs serve highly specific purposes in statutory language: suasive verbs mandating or giving direction (or excusing from responsibility) (excerpt 4) and verbs of causation giving permission to act (excerpt 5).

(4) **Excerpt 4**: Zoo animals loaned pursuant to this section are not deemed to be surplus property, and **no** motion **is required** to enter into an agreement for the loaning of zoo animals. (S.D. § 6-13-16).

(5) **Excerpt 5**: If the tax collector fails to respond at any step in the process under this section within the prescribed period of time, then the governing body **shall be permitted** to remove the tax collector from office as provided in paragraph V. (N.H. § 41:40).

Excerpt 5 also demonstrates the use of modal *shall* (common in legislation) and the passive voice, the latter of which is another characteristic feature of literate varieties such as formal academic writing. The passive voice has historically been given lots of attention in the conversation surrounding readability of texts and is frequently targeted in text simplification. Two forms of the passive voice are key with large effect sizes in statutory law (*by* passives and short passives), corroborating past findings by Williams (2004) and Bulatović (2013). While short passives are typically favored when the agent is unknown, as is common in academic writing (Biber *et al.* 2002: 168), they seem instead to be favored in statutory law for the purpose of inclusiveness. In many cases, leaving out the agent necessarily implies ‘everybody’ or ‘anybody’ who commits an act, which, importantly, makes it clear that all citizens of that state are subject to that particular law, and the legal consequences should they not act in accordance with it. On the other hand, the passive + *by* construction is used in statutory language for the opposite purpose: to indicate exactly who has the authority or power to act. Note the passive constructions in excerpts 6 and 7, which are used in two different ways: 1) to indicate that an action applies to everyone, and 2) to give a person or entity authority.

(6) **Excerpt 6**: The official flag of the state shall **be displayed** with the flag of the United States only from sunrise to sunset, or between the hours **designated by** proper authority. However, the flag may **be displayed** after sunset upon special occasions when it is desired to produce a patriotic effect. (A.K. § 44.09.030).

(7) **Excerpt 7**: If the date of the special election **conducted** pursuant to § 12-11-1.1 requires that absentee ballots **cast by** absent uniformed services voters or overseas voters arriving after election day be counted as **required by** 2 USC Chapter 1 § 8 as of January 1, 2008, these absentee ballots shall **be processed and counted by** the provisional ballot counting board. (S.D. § 12-11-2.1).

Results of the key feature analysis show that statutory language exhibits the use of both clausal and phrasal features, confirming the findings by Goźdź-Roszkowski (2011). Notable phrasal features that are key in statutory law include prepositions and phrasal conjunctions, which serve the purpose of providing as much detail as possible in the description of the person, context, or situation in which a law applies. Prepositions, which signal embedded prepositional phrases and phrasal verbs, have the largest effect size of any feature in the corpus of statutory law (d=1.84). In statutory language, they function predominantly to provide qualifying details in order to narrow the identity or scope of the noun that they modify. The use of this contributes to the dense packaging of referential information, as they are more compact than clausal postmodifiers (Biber *et al.* 1999: 607). Excerpt 8, below, comprises a single sentence with ten prepositions (in bold), which together function to provide an operational definition of a term. One of these prepositions belongs to a single complex prepositional phrase (*with respect to*), one is a part of a prepositional verb (*deal with*), and six prepositions head a prepositional phrase. These prepositional phrases come in a variety of forms, including genitive/postmodifying (e.g., *law [of this state]*), and adverbial (e.g., *property [within this state]*).

(8) **Excerpt 8**: (a) "Charitable trust" means any fiduciary relationship **with** respect **to** property arising under the law [**of** this state] or [**of** another jurisdiction] as a result [**of** a manifestation] [**of** intention] to create it and subjecting the person by whom the property is held to fiduciary duties to deal **with** the property [**within** this state] for any charitable, nonprofit, educational, or community purpose. (N.H. § 7:21).

Phrasal embedding also appears in the form of phrasal coordinating conjunctions, which, along with the use of causal coordinating conjunctions, contribute to the long, drawn-out sentences that are packed with information and tend to make sentences hard to follow (Tiersma 1999). Phrasal and clausal coordinating conjunctions are both key in the corpus of statutory law with large effect sizes over 1.0 (d=1.64; d=1.10). Phrasal coordinators in statutory language are most often used to directly identify a highly specific list of individuals, entities, or objects that the statute applies to. See, for instance, excerpt 9, which lists a set of qualifying items (*labor*, *material*, or *rental*

equipment), that must be furnished by the person in order for the statute to apply to them. The length of excerpt 9, which contains a total of ten phrasal and clausal conjunctions, is attributed to the thorough description of the circumstances under which an individual has the right to sue. This results in the subject (a person) being separated from the corresponding verb phrase *has the right to*, by a string of embedded clauses and phrases, including seven phrasal and clausal conjunctions. This format is not uncommon, as describing the characteristics of the subject that a statute pertains to, or the context in which the statute takes effect, is an important characteristic of statutory language.

> (9) **Excerpt 9**: (c) **A person** *who has furnished labor,* material, or rental equipment to a bonded contractor or his subcontractors for the work specified in the contract, and *who has not been paid in full* for it before the expiration of a period of ninety days after the day on which the last of the labor was done or performed by the person or material or rental equipment was furnished or supplied by the person for which the claim is made, **has the right to sue** on the payment bond for the amount, or the balance of it, unpaid at the time of institution of the suit and to prosecute the action for the sum or sums justly due the person. (S.C. § 11-35-3030).

The use of multiple phrasal and clausal coordinators in quick succession to one another can result in confusion, as it can be easy to mistake one type of conjunction for the other. For example, in excerpt 9, if one reads: *the last of the labor was done or performed by the person or material or rental equipment (...)*, it is easy to mistake the second clausal conjunction (*person or material…*) for a phrasal conjunction. This is resolved semantically in the clause, but is still challenging to process in real time as the reader looks for a conclusion to the long sentence in the form of a phrasal conjunction, and is instead introduced to yet another clause.

Conditional subordinating conjunctions (e.g., *if, unless*) are frequently associated with statutory language for the same reasons mentioned above: they contribute to the specification of the conditions under which authorizations, mandates, or prohibitions take effect, or do not take effect. Excerpt 10, below, includes five conditional statements in a list format, each moving further from the initial clause that the conditional statement is dependent upon for meaning. Because of this, and because conditional subordinating clauses have flexibility in their syntactic position (i.e., beginning, medial, final), the conditional statement can start to read as though it occupies the beginning syntactic position. This is particularly problematic for readability when the sentence

potentially reads more smoothly with the conditional statement in a different syntactic position from which it appears.

(10) **Excerpt 10**: (j) Upon conviction by a court of a person of an offense described in (a)(7) of this section, the department shall disqualify that person from driving a commercial motor vehicle for the following periods:

1. **if the person has not been previously convicted of violating** an out-of-service order, not less than 180 days;
2. **if the person has been previously convicted once** of violating an out-of-service order, not less than two years;
3. **if the person has been previously convicted more than once** of violating an out-of-service order, not less than three years;
4. **if the person operates a commercial motor vehicle transporting hazardous materials** or a vehicle designed to transport 16 or more passengers, including the driver, in violation of an out-of-service order, not less than 180 days;
5. **if the person has been previously convicted of operating a commercial motor vehicle transporting hazardous materials** or a vehicle designed to transport 16 or more passengers, including the driver, in violation of an out-of-service order two or more times in separate incidents within a 10-year period, not less than three years. (A.K. § 28.33.140).

While highly clausal language is frequently associated with decreased readability of statutes, the key feature analysis in the present study revealed that the preference for the type of clause may be what distinguishes statutory language from other forms of popular language. In particular, there was a difference in the distribution of finite and non-finite causal constructions: statutory language uses markedly more non-finite clauses relative to popular written language, and markedly fewer finite clauses. All key clausal constructions in the PWL were finite, including *that* relative clauses, *that* complement clauses and *wh-* subject position relative clauses, and nearly all key clausal constructions in the statutory law corpus were non-finite, including post-nominal *-ing* and *-ed* clauses, and two types of *to-* complement clauses. The exception to this pattern was the *wh*-relative clause with the pronoun in the object position (both with prepositional fronting and without) which appeared alongside the non-finite constructions in statutory law. This distribution should be interpreted with caution, however, as only five of the nine clausal constructions meet the Cohen's *d* threshold for 'key' (>-/+0.20). The distribution of non-finite and finite clauses is represented in Figure 2.

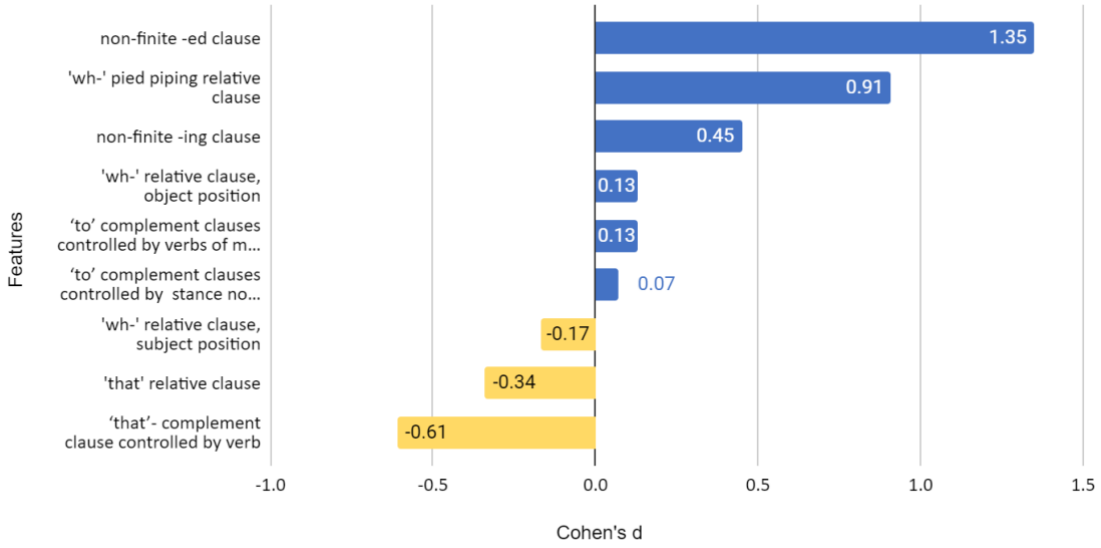

Figure 2: Distribution of finite and non-finite clause constructions

Non-finite clauses, which are favored in statutory law, are more compact and less explicit than finite clauses and often lack an explicit subject or subordinator (Biber *et al.* 1999: 198). This lack of explicitness is directly related to the condensed nature of statutes, as the drafters attempt to pack as much information as possible into a small space. The condensed nature of the non-finite *-ing* and *-ed* clauses can be seen in excerpts 11 and 12.

(11) **Excerpt 11**: A mutual bank may, with the approval of the department, establish and operate branches inside the state. Before **approving the establishmen**t and operation of a branch xoffice, the department shall make the findings required before the granting of a charter to a mutual bank with respect to the **branch proposed**. (A.K. § 06.15.290).

(12) **Excerpt 12**: The governing body, within sixty days after the filing of any such delinquent list, shall examine such list and, on **being** satisfied that any of the taxes **so listed** are not collectible, it shall, by resolution, release the collector from the collection thereof and order **the same canceled.** (N.J. § 54:4-91.2).

In contrast, finite clauses such as *that* complement clauses controlled by verbs and *that* relative clauses were markedly less common in the statutory law corpus. *That* complement clauses controlled by verbs, which have a medium effect size of d = -0.61 in the PWL corpus, are often used in reported or quoted speech, which is highly common in texts narrating past events or recalling conversations, as shown in excerpt 13.

(13) **Excerpt 13**: Encyclopedia article. Speaking about Niall Horan, who we made his boyfriend when we created Nithan Syran, Nat Han said: "I love Niall...Can I just **say that** Niall is one of the nicest lads you'll ever. (http://www.sugarscape.com/main-topics/celebrities/784664/exclusive-nathan-sykes-wanted-and-one-direction-being-lovers)

That relative clauses (key in the PWL corpus with a small effect size of d = -0.34) are used in the post-modification of a noun phrase in either the restrictive form (to establish a reference) or non-restrictive form (to provide additional information about the antecedent, not required for identification). These two forms can be seen in the encyclopedia excerpt (14) below.

(14) **Excerpt 14**: Encyclopedia article. While in common parlance anything **that** attempts to provide an explanation for a cause can be dubbed a "theory", a scientific theory has a much more specific meaning. Scientific theory is far more than just a casual conjecture or some Joe's guesswork. A theory in this context is a well-substantiated explanation for a series of facts and observations **that** is testable and can be used to predict future observations. (*Scientific Theory*: https://rationalwiki.org)

While these functions are also important in statutory language, statutes appear to favor *wh-* relativizers. This is consistent with the findings of Biber *et al.* (1999: 611), who found a preference for *wh-* relativizers in formal academic prose over *that* relativizers (twice the frequency of occurrence).

The notable exception to the finite/non-finite split between the two registers is the pied-piping *wh-* relative clause construction (d=0.91). In this construction, there is a preposition located at the beginning of the clause preceding the relative object pronoun, resulting in classic formal phrases such as *to whom*, *for which*, and *in which*. The prepositional fronting does not serve any immediate, unique function in statutory law, but is instead considered stylistic and highly characteristic of statutory language. In excerpt 15, below, a single sentence holds four instances of this construction. Excerpt 16 also makes use of four pied-piping constructions embedded within an even longer sentence, which has been truncated in order to conserve space. It should also be noted that the pied-piping construction is frequently passive, a characteristic again shared with academic prose as found by Biber *et al.* (1999), who noted that object position relative clauses in particular are frequently found alongside the passive voice.

(15) **Excerpt 15**: A remote claimant has a right of action on the payment bond only upon giving written notice to the contractor within ninety days from the date **on which the person did or performed the last of the labor** or furnished or supplied the last of the material or rental equipment **upon which the claim is made**, stating with substantial accuracy the amount claimed as unpaid and the name of the party **to whom the material or rental equipment was furnished or supplied** or **for whom the labor was done or performed**. (S.C. § 11-35-3030).

(16) **Excerpt 16**: When any owner, tenant or subtenant of a lot or lots or tract of land shall file in any court of competent jurisdiction within the county **in which said lot or lots or tract of land may be situated**, his or her affidavit, or the affidavit of any other creditable person for them, stating that from knowledge, information or belief the party or parties owning, controlling or working the adjoining lot or lots or tract of land, and **upon which said party or parties are sinking shafts**, mining, excavating and running drifts, and that said drifts, **in which said parties are digging, mining and excavating** any mineral ore or veins of coal, extend beyond the lines and boundaries of said lot or lots or tract of land owned, controlled or worked by them, and have entered into and upon the premises of the party or parties making said affidavit, or **for whom said affidavit is made**, the judge of such court shall issue his or her written order [...]. (K.S. § 49-109).

An important finding of the present study is that clausal embedding as a whole cannot necessarily be considered highly characteristic of statutory language relative to other forms of written language. This is based on two findings: 1) different types of clausal constructions appeared key in both corpora, and 2) several constructions had effect sizes approaching zero, indicating similar use in statutory language and popular written language. The latter finding is demonstrated in the two excerpts below, which demonstrate similar use of a variety of clausal features in statutes and written popular language.

(17) **Excerpt 17**: *Wh-* relative clause, subject position (*who*); *Wh-* relative clause, object position (*which*)
(c) Any member **who** is aggrieved by a denial of benefits to be provided under this section may appeal the denial in accordance with regulations of the department of health, **which** have been promulgated pursuant to chapter 17.12 of title 23. (R.I. § 27-30-1).

(18) **Excerpt 18**: Encyclopedia article. *Wh-* relative clause, subject position.
The remainder of your companions in the following order of priority, minus whoever is already included in your active party and those **who** have sided against you before this point [...]. (https://dragonage.fandom.com/wiki/The_Last_Straw)

This suggests that the *type* of clausal construction may matter quite a bit for readability. For this reason, it seems that the discussion surrounding clausal constructions that are particularly problematic should focus more narrowly on constructions that are markedly less common in popular written language, and particularly characteristic of statutory language, such as the *wh-* pied-piping construction and the condensed non-finite *-ed* and *-ing* clauses.

6. Conclusions and future research

This study has provided a large-scale, detailed description of what the register of statutory language looks like and, in particular, how it differs from language that the lay person is exposed to on an everyday basis. It is important that we continue to make these comparative analyses when we attempt to describe statutory language so that we understand not just how frequently a feature appears in register, but how *characteristic* it is of that register. For example, Hiltunen (2012) reported that around a quarter of subordinating clauses in legislation are adverbial, and while this seems like a large proportion, it may not paint the full picture of the use of this feature. In the present study, the feature 'other adverbial subordinating clauses' actually had a medium effect size in the PWL corpus (see Table 5), meaning that it is markedly *less* frequent in statutes compared to other forms of popular written language.

This study has also showed us that we need to be looking at clausal embedding at a more fine-grained level as opposed to making blanket statements about the challenges that it poses for readability. The present study has demonstrated that several types of finite clauses, for example, are in fact key in the PWL corpus, or not key at all (exceedingly small effect sizes, under +/- .20).

Future research of this kind would benefit from a more detailed analysis of clausal embedding, with a specific focus on adverbial clauses and centrally-embedded clauses, which both Charrow and Charrow (1979) and Bhatia (1993) argue are highly characteristic of legislative language and problematic for readability. Future research may also expand on this information to examine readability from the reader's perspective. It is hoped that this study may provide a constructive path forward in addressing lack of readability in legislative texts, both by demonstrating the use of

empirical methods to identify differences between statutory and popular language, and identifying features that may be less familiar to the lay person.

REFERENCES

Alasmary, Abdullah. 2019. Lexical bundles in contract law texts: A corpus-based exploration and implications for legal education. *International Journal of English Linguistics* 9/2: 244–257.

Andersson, Dan. 2007. *Deontic Modal Verbs in EU Legislation: A Comparative Study of Documents in Four Germanic Languages*. Stockholm: University of Stockholm dissertation.

Bednarek, Monika and Helen Caple. 2014. Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society* 25/2: 135–158.

Bhatia, Vijay Kumar. 1983. Simplification v. easification: The case of legal texts. *Applied Linguistics* 4/1: 42–54.

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

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Language in Contrast* 14/1: 7–34.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style.* Cambridge: Cambridge University Press.

Biber, Douglas and Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics* 44/2: 95–137.

Biber, Douglas and Jesse Egbert. 2018. *Register Variation Online*. Cambridge: Cambridge University Press.

Biber, Douglas and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing.* Cambridge: Cambridge University Press.

Biber, Douglas and Bethany Gray. 2019. Are law reports an 'agile' or an 'uptight' register? Tracking patterns of historical change in the use of colloquial and complexity features. In Teresa Fanego and Paula Rodríguez-Puente eds. *Corpus-based Research on Variation in English Legal Discourse*. Amsterdam: John Benjamins, 147–170.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd and Marie Helt. 2002. Speaking and writing in the university: A multidimensional comparison. *Tesol Quarterly* 36/1: 9–48.

Biber, Douglas, Stig Johansson, Geofrey Leech, Susan Conrad and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Biel, Lucja. 2009. Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In Carmen Heine and Jan Engberg eds. *Reconceptualizing LSP. Proceedings of the XVII European LSP Symposium* Aarhus: Aarhus University, 1–15.

Biel, Lucja. 2014. Phraseology in legal translation: A corpus-based analysis of textual mapping in EU law. In Le Cheng, King Kui Sin and Anne Wagner eds. *The Ashgate Handbook of Legal Translation*. London: Routledge, 178–192.

Biel, Lucja. 2017. Lexical bundles in EU law: The impact of translation process on the patterning of legal language. In Stanislwaw Goźdź-Roszkowski and Gianluca Pontrandolfo eds. *Phraseology in Legal and Institutional Settings: A Corpus-Based Interdisciplinary Perspective.* London: Routledge, 10–26.

Breeze, Ruth. 2013. Lexical bundles across four legal genres. *International Journal of Corpus Linguistics* 18/2: 229–253.

Bulatović, Vesna. 2013. Legal language: The passive voice myth. *ESP Today* 1/1: 93–112.

Caliendo, Giuditta, Gabriella Di Martino and Marco Venuti. 2005. Language and discourse features of EU secondary legislation. In Anna Duszak and Guiseppina Cortese eds. *Identity, Community, Discourse: English in Intercultural Settings*. Bern: Peter Lang, 381–404.

Charrow, Robert P. and Veda Charrow. 1979. Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review* 79: 1306–1374.

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Davies, Mark. 2012. *Oppositions and Ideology in News Discourse*. London: Bloomsbury Academic Press.

Davies, Mark. 2013. *Corpus of Global Web-Based English*. https://corpus.byu.edu/glowbe/

Egbert, Jesse and Margaret Wood. (Under review). *Constructing and Designing a Specialized Corpus of Statutory Law* (CorUSSS).

Foley, Roger. 2002. Legislative language in the EU: The Crucible. *International Journal for the Semiotics of Law* 15/4: 361–374.

Fowler, Roger. 2013. *Language in the News: Discourse and Ideology in the Press*. London: Routledge.

Gibová, Klaudia. 2011. On modality in EU institutional-legal documents. In Alena Kačmárová eds. *English Matters II: A Collection of Papers by the Institute of British and American Studies Faculty*. Prešov: University of Prešov, 6–12.

Goźdź-Roszkowski, Stanislaw. 2011. *Patterns of Linguistic Variation in American Legal English: A Corpus-based Study*. Bern: Peter Lang.

Hiltunen, Risto. 2012. The grammar and structure of legal texts. In Lawrence M. Solan and Peter M. Tiersma eds. *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 39–51.

Jablonkai, Réka. 2010. English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes* 29/4: 253–267.

Özyildirim, Işil. 2011. A comparative register perspective on Turkish legislative language. *Law Review* 1/1: 79–94.

Pontrandolfo, Gianluca. 2015. Investigating judicial phraseology with COSPE: A contrastive corpus-based study. In Claudio Fantinuoli and Federico Zanettin eds. *New Directions in Corpus-based Translation Studies*. Berlin: Language Sciences Press, 137–160.

Rodríguez-Puente, Paula. 2019. Interpersonality in legal written discourse: A diachronic analysis of personal pronouns in law reports, 1535-present. In Teresa Fanego and Paula Rodríguez-Puente eds. *Corpus-based Research on Variation in English Legal Discourse*. Amsterdam: John Benjamins, 171–199.

Scollon, Ron. 2014. *Mediated Discourse as Social Interaction: A Study of News Discourse*. London: Routledge.

Seracini, Francesca. 2020. *The Translation of European Union Legislation: A Corpus-based Study of Norms and Modality*. Milan: LED Edizioni Universitarie.

Staples, Shelley, Jesse Egbert, Douglas Biber and Bethany Gray. 2016. Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication* 33: 149–183.

Sun, Yuxiu and Le Cheng. 2017. Linguistic variation and legal representation in legislative discourse: A corpus-based multi-dimensional study. *International Journal of Legal Discourse* 2/2: 315–339.

Tapia, Ana M. Gates and Douglas Biber. 2014. Lexico-grammatical stance in Spanish news reportage: Socio-political influences on *que*-complement clauses and adverbials in Ecuadorian broadsheets. *Spanish Journal of Applied Linguistics* 27/1: 208–237.

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

Trebits, Anna. 2009. The most frequent phrasal verbs in English language EU documents: A corpus-based analysis and its implications. *System* 37/3: 470–481.

Williams, Christopher. 2004. Legal English and plain language: An introduction. *ESP across Cultures* 1/1: 111–124.

Williams, Christopher. 2007. *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts*. Bern: Peter Lang.

Williams, Christopher. 2013. Changes in the verb phrase in legislative language in English. In Bas Aarts, Jo Close and Sean Wallis eds. *The Verb phrase in English: Investigating Recent Language Change with Corpora*, 353–371.

Xie, Qin. 2018. Critical discourse analysis of news discourse. *Theory and Practice in Language Studies* 8/4: 399–403.

Yana, Dewi. 2015. The lexico grammatical features of the political register analysis in the editorial of the Jakarta Post Newspaper. *ANGLO-SAXON* 6/2: 15–23.

Corresponding author
Margaret Wood
Northern Arizona University
Department of English
705 S Beaver St
Flagstaff, AZ 86001
United States
E-mail: mkw57@nau.edu

received: December 2021
accepted: June 2022

APPENDIX 1: Statutory law and popular written language varieties situational characteristics

Register	Participants	Relationship among participants	Production circumstances	Setting	Purposes	Topic
State Codes	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Unequal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit	Varied (divorce, wills, personal injury, welfare, crime, real estate)
News Reports	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Varied (politics, economy, entertainment, business, health)
Sports Reports	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Sports
Encyclopedia Articles	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Varied
Historical Articles	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Varied