An investigation of the lexico-grammatical profile of English legal- lay language

Lucia Busso

Aston Institute for Forensic Linguistics, Aston University, UK

https://doi.org/10.21747/21833745/lanlaw/9_1a7

Abstract. *The article presents a study on the lexico-grammar of the genre of English legal-lay language (Tiersma 1999), using the English subcorpus of the CorIELLS corpus (Busso forthcoming). The study explores four grammatical constructions (in Goldberg 2006's Construction Grammar sense): nominalisations heading prepositional phrase attachments, modal verb constructions, participial reduced relative constructions, and passive constructions. Specifically, we use collostructional analysis (Stefanowitsch 2013), followed by a vocabulary analysis using English core vocabulary as a reference (Brezina and Gablasova 2015), and a comparative frequency analysis with corpora of legal language and general-domain written prose. Results of this first part of the study foreground how legal-lay language is quantitatively different from both neighbouring genres, suggesting that it might be considered a "blended" genre. We further explore the data in terms of accessibility for speakers, using readability metrics and a survey on English participants. Both methods show that legal-lay language is at an intermediate level of complexity between legal jargon and general-domain prose; however, we further note that readability metrics generally underestimate speakers' ability to comprehend legal-lay language.*

Keywords: *Construction Grammar, Legal lay language, Italian, English, Quantitative corpus linguistics.*

Resumo. *O artigo apresenta um estudo sobre a léxico-gramática do género linguagem jurídica para leigos (Tiersma 1999), utilizando o subcorpus inglês do corpus CorIELLS (Busso forthcoming). O estudo explora quatro construções gramaticais (no sentido da Gramática da Construção de Goldberg 2006): nominalizações que regem sintagmas preposicionais, construções com verbos modais, construções relativas restritivas participiais e construções passivas. Especificamente, recorremos à análise colostrucional (Stefanowitsch 2013), seguida de uma análise de vocabulário utilizando o vocabulário principal do inglês como referência (Brezina and Gablasova 2015), e uma análise de frequência comparativa com corpora de linguagem jurídica e textos escritos de linguagem* *geral. Os resultados desta primeira parte do estudo destacam como a linguagem jurídica para leigos é substancialmente diferente dos dois géneros que lhe estão próximos, o que sugere tratar-se de um género "híbrido". Exploramos, ainda, os dados em termos de acessibilidade para os falantes, recorrendo a métricas de compreensibilidade e a um inquérito realizado junto de participantes ingleses. Os dois métodos mostram que a linguagem jurídica destinada a leigos se encontra num nível intermédio de complexidade entre o jargão jurídico e a linguagem geral; contudo, notamos ainda que as métricas de legibilidade subestimam, habitualmente, a capacidade de os falantes compreenderem a linguagem jurídica para leigos.*

Palavras-chave: *Gramática construtiva, Linguagem jurídica para leigos, Italiano, Inglês, Linguística de corpus quantitativa.*

Introduction

Comprehension and readability of legal documents – especially if aimed at non-specialists – has been at the centre of the debate in both applied linguistics and legal studies (Tiersma 1999; Frade 2007; Haapio 2011). Particularly, many scholars have advocated a clearer, plainer language in the drafting of legal texts for the lay public (Charrow and Charrow 1979; Schiess 2007).

There is often a difficult trade-off to manage when deciding the level of linguistic complexity to embed in a LLL text, which has to obtain both legal precision and linguistic clarity. In fact, writing a legal document that is at the same time clear and understandable and respects the intricacies of the law is not an easy task (Ződi 2019). However, while a certain level of complexity due to the topic is generally considered vital to reduce vagueness as much as possible (Gotti 2014), a lack of comprehensibility leads to linguistic and legal problems alike (Haapio 2011; Conklin *et al.* 2019).

The need for more comprehensible language in legal settings has been present among English scholars for a long time and resulted in the *Plain Language* movement (Bhatia 1983; Adler 2012) , the most prominent example of interdisciplinary effort to simplify access to complex texts. As Adler (2012: 3) specifies,

> " '[p]lain language' means language and design that presents information to its intended readers in a way that allows them, with *as little effort as the complexity of the subject permits*, to understand the writer's meaning and to use the document."

In the UK, the *Plain English Campaign* (founded in 1979) has been "campaigning against gobbledygook, jargon and misleading public information." (Plain English Campaign website).

In general, scholars advocating for a simplification of legalese argue that syntactic, semantic, and pragmatic complexity hinder comprehension for the lay reader. This issue is crucially relevant in contemporary societies, where different (often binding) legal documents regulate many parts of every-day life. The online world in fact constantly exposes us to legal texts – which require from the user a basic understanding of legal concepts for a variety of purposes (e.g., the terms and conditions of websites, legal notices of online banking services, etc.). The overwhelming importance of legal-lay language is perfectly exemplified by the Cambridge Analytica scandal, in which legally binding terms and conditions of an app used by Facebook stated that they were harvesting data from users who authorised it and their friends. The app then would transfer the data to the political consultancy Cambridge Analytica which could assemble psychological profiles of voters based on their online presence (Romm 2018).

Terms and conditions are the most used example of legal-lay language; however, it is here argued that the term expands wider than that, including all types of texts with legal content but aimed at a non-specialist audience (Tiersma 1999; Williams 2010; Busso 2022, forthcoming). Example 1 and 2 represent two concordances of the word "contract" extracted respectively from the corpora of legal-lay and specialist legal language used in this article.

1. A contract for the provision of an account with the functions described in these Terms and Conditions is concluded when we confirm that we have set up an account for you either via e-mail or through a message delivered through the App.
2. The Court shall have jurisdiction to give judgment pursuant to any arbitration clause contained in a contract concluded by or on behalf of the Community, whether that contract be governed by public or private law.

Specifically, it is here argued that specialised legal jargon and legal-lay language (henceforth: LLL) can be considered – at least in some respects – different. In fact, the inaccessible nature of legal texts has been widely studied (Chovanec 2013). Complex and highly specialised syntax and lexicon are the most noticeable features of this genre, playing an almost 'ritualistic' role in identifying it (Coulthard and Johnson 2007: 37). But while specialist legal language remains principally used by professionals with years of legal education, LLL has instead the specific aim to be read and – more critically – understood by lay readers.

As mentioned, legal language has been extensively researched by linguists. Most recently, many scholars have started to use computational models to analyse the genre (Hamann *et al.* 2016; Fanego and Rodríguez-Puente 2019; Van Boom *et al.* 2016; Frankenreiter and Livermore 2020). However, the linguistic analysis of LLL as a separate textual type is still an under researched area (Lintao and Madrunio 2015; Conklin *et al.* 2019). The present contribution aims at filling this gap by providing an exploratory analysis of a corpus of LLL in English, following a similar procedure to the study outlined in Busso (2022) for Italian. Specifically, combining evidence from quantitative text-based and experimental methods, this article addresses the following research questions, which focus on different level of linguistic analysis:

I. How specialised is the LLL lexicon? (lexico-grammatical level)
II. Does LLL exhibit linguistic features that are measurably different from specialist legal jargon and general-domain written language alike? (syntactic-semantic level)
III. How comprehensible is LLL with respect to legal and general-domain written language? (semantic-pragmatic level)

Construction Grammar as a reference framework for corpus-based analysis

Construction Grammar, a family of linguistic theories advocating a Usage-Based model of language, understands language as composed of complex units called *constructions*

(Goldberg 2006, 2019). Constructions are conceptually cognate to the Saussurean notion of *sign* as a "two-sided psychological entity" (Saussure 1916: 63) that combines a particular form, i.e., the 'signifier' (or 'signifiant'), with a particular meaning, i.e., the 'signified' (or 'signifié'). Crucially, Construction Grammar extends the idea of arbitrary form-meaning pairings to all levels of grammatical description – from lexical items, to abstract phrasal patterns.[1] Since constructions at the lexical level are not ontologically different from abstract grammatical constructions, Construction Grammar does not see syntax and lexicon as qualitatively different as in rule-based models of language (Pollard and Sag 1994). All constructions – from lexical items to fully schematic syntactic structures – are included in the *constructicon* of a language, i.e., the full inventory of constructions. In other words, constructions differ among themselves only in terms of length, complexity, or level of schematicity.

Since Construction Grammar is part of the constellation of Usage-based models – i.e., models that argue that knowledge of usage is inseparable from grammar (Bybee 2015) – observational data such as corpus data play a crucial role in many studies that adopt such a framework (Gries 2013; Hilpert 2013). Furthermore, it has been argued that Construction Grammar can prove to be useful for the analysis of genre (Groom 2019). A constructionist approach hence offers tools for an approach to text analysis that allows for a cohesive and unified account of features at different levels of linguistic complexity.

The present study uses corpus data aligning itself to general constructionist tenets. That is, the study foregrounds usage of form-meaning patterns at different levels of abstraction, and analyses their structure, function, and frequency. While the study of morpho-syntactic patterns in legal and bureaucratic language is by no means uncommon in the literature (Goźdź Roszkowski and Pontrandolfo 2015; Goźdź-Roszkowski and Pontrandolfo 2017), there is no study in the literature – to the best of the author's knowledge – that explicitly uses Construction Grammar as a means to explore the linguistic structure of legal or legal-lay language. Moreover, analyses of phraseology in legal contexts are mainly qualitative in nature, while the present work employs quantitative methods.

The next sections will provide an in-depth description of the data and of the various methods of analysis used: collostructional analysis, lexical analysis, and contrastive frequency analyses comparing LLL to legal and written prose corpora. The last section draws general conclusions from the analyses performed.

Data: the *CorIELLS* corpus

As data, the study employs *CorIELLS* (CORpus of Italian and English Legal-lay textS), a specialised bilingual corpus of LLL in Italian and English (Busso 2022)[2]. In line with our working definition of LLL (see Introduction section), different textual types were included in the corpus. Particularly, the types of document selected follow two general criteria for inclusion. They are all: (a) freely available online, to approximate the types of LLL people are exposed to on the Internet, and (b) varied, to obtain a sample as representative as possible for the genre in question (Biber 1993; Almut 2010). The final selection includes four major categories of document:

- TERMS AND CONDITIONS AND/OR TERMS OF USE OF WEBSITES. 45 in total or for each country websites were manually selected from the Alexa list of the

500 most visited websites in Italy and the UK in 2019; only web services with legal notices in both languages were included.
- EUROPEAN LEGISLATION SUMMARIES. These texts are "short, easy-to-understand explanations of the main legal acts passed by the EU and intended for a general, non-specialist audience" *(EUR-lex* website).[3] A selection of texts was collected from the official website *EUR-lex* in both their Italian and English versions. 247 summaries per language (all summaries from 2019 and 2020).
- BANK CONTRACTS. Freely accessible legal documents for standard current accounts were selected from 15 banks in Italy and the UK.
- UTILITIES. Standard contract terms for 5 energy suppliers, 5 Wi-Fi suppliers, and 5 pay-by-the-month phones in Italy and the UK were selected.

The documents were semi-automatically retrieved, cleaned, and downloaded using the web scraping *Bootcat* toolkit (Baroni and Bernardini 2004). Size of the corpus amounts to 1.85M words. Composition of the general corpus can be seen in Table 1.

Document type	**Number of texts**	**English subcorpus (800K words)**	**Italian subcorpus (1M words)**
Bank contracts	15 per language	19%	27%
Utilities contracts	15 per language	25%	23%
Terms and conditions	45 per language	34%	27%
EurLex summaries	247 per language	22%	23%

Table 1. Composition of CorIELLS and of its English and Italian subcorpora

In this article, only the English subcorpus will be analysed. For a similar study using the Italian subcorpus of CorIELLS, see Busso (2022).

Construction selection

Lexical bundles and grammatical patterns are a common object of study in the analysis of legal and bureaucratic language (Goźdź Roszkowski and Pontrandolfo 2015; Goźdź-Roszkowski and Pontrandolfo 2017; Yunus and Ab Rashid 2016). However, as mentioned in the Introduction, this paper takes the analysis of phraseology a step further, considering lexico-grammatical patterns as constructions, i.e., linguistic units.

Four constructions were selected for two theoretically motivated reasons. Firstly, constructions were selected at different levels of abstraction to obtain a balanced representation of the lexico-grammatical nature of the corpus. Secondly, the selection was carried out capitalizing on previous research on legal and bureaucratic texts; only constructions unanimously recognized by the literature as highly characteristic of legal language and LLL were selected (Williams 2004; Chovanec 2013; Haigh 2013).

i. Lexical/phrase level: Nominalizations heading prepositional chains (henceforth: NOM_PP). Nominalizations are lexical constructions broadly defined lexically as the "process via which a prototypical verbal clause (…) is converted into a noun phrase" (Givón 1993: 287) . They have been long recognised as being "overwhelmingly used in legislative provisions" (Bhatia 1993: 148). This type of construction is especially used instead of verb phrases (VP), which are usually scarce in English legal texts (Williams 2013). That is, events are preferentially encoded through deverbal nominalizations, typically embedded in long PP-attachment chains, as in example 1.[4]

   1.
   2.
   3. Mandatory collective management of rights for retransmissions of radio and television programmes by means other than cable.

ii. Phrase Level: Modal verbs (henceforth: MOD). Phraseological patterns Vmod + V composed by a modal verb and any finite or non-finite form of a verb, as shown in example 4 below. MOD are generally understood as ‘grammaticized constructions’ (Langacker 2013: 14; see among others Cappelle and Depraetere 2016; Torres–Martínez 2019). The literature has long recognized modality as one of the distinguishing features of legal and bureaucratic language (Tiersma 1999; Aher 2013).

   1.
   2.
   3.
   4. We must be satisfied of your identity and can refuse instructions if we doubt your identity.

iii. Phrase/Clause Level: Reduced participial relative clause (henceforth: PART). These constructions contain a present (or past) participle that ‘replaces’ a relative pronoun and main verb (Quirk *et al.* 1985). Present participial constructions are typical of the morpho-syntax of legal English (Janigová 2008).

   1.
   2.
   3.
   4.
   5. The ‘application publisher’ means the entity licensing the application to you as identified in the Store.

iv. Discourse level: Passive constructions (henceforth: PASS) (Jaeggli 1986)Passive constructions are a distinctive feature of legal and legal-lay texts (Bulatović 2013), often used to omit the agent of the sentence, as in example 4 below. It has been claimed in the literature that an excessive use of passives leads to highly cognitively demanding texts (Yokoyama *et al.* 2006).

   1.
   2.
   3.
   4.
   5.
   6. Payments (. . .) will be sent on the next working day.

The chosen constructions are used for all following analyses. To retrieve all instances of them in the corpus, general CQL queries were carried out on *SketchEngine* (Kilgarriff *et al.* 2004).[5]

Collostructional analysis

Collostructions vs collocations

Collostructional analysis (Stefanowitsch, 2013) is a family of quantitative methods that measure the statistical preference or dispreference (in terms of association strength) that words exhibit to constructions. It is an extension of traditional collocational analysis using Construction Grammar tenets; the term *collostruction* itself is in fact a blend of the two words 'collocation' and 'construction'. It significantly differs from traditional collocation methods since it does not measure the association of words to other words, but of words to syntactic patterns.

Since meaning of abstract constructions is understood to emerge from the meaning of its fillers, collostructional analysis contributes to the identification of the meaning range of constructions. In other words, using collostructional analysis helps to discover how a construction is used. Words that are found to be significantly attracted to the analysed constructions are called *collexemes*.

Collostructional analysis is composed of three types of methods: *simple, distinctive*, and *covarying* collexeme analysis. In this paper, simple and covarying collexeme analysis will be used. Simple collexeme analysis (Stefanowitsch and Gries 2003)(henceforth: SC) is the clearest reinterpretation of collocational analysis in a grammatical perspective. It measures the statistical co-occurrence relation of a lemma to a slot in a construction (typically an argument structure construction). Co-varying collexeme analysis (Stefanowitsch and Gries 2005)(henceforth: CC) is used instead to quantify the association of lemmas in one slot to lemmas in another slot of a single construction.

These two methodologies are employed on the four constructions selected (see ***Construction selection*** subsection). More specifically, simple collexeme analysis was carried out for NOM_PP, PART, and PASS. The investigated slots are respectively the deverbal noun, the present participle, and the main verb. Covarying collexeme analysis is instead performed for MOD, retrieving association strength for modal and main verb. Both analyses are conducted using the R package *collostructions* (Flach 2018).

Simple collexeme (SC) and Covarying collexeme (CC) analysis

To perform SC, CQL queries of the general constructions were performed on the web corpus tool *SketchEngine* and a frequency list of all the lemmas in the fillers under consideration was extracted. Data were then manually checked and cleaned from noise. The analysis was carried out on all lemmas occurring with a frequency equal to or higher than 5 for NOM_PP and PART, and on the first 100 occurrences for PASS. The final dataset consists of 50 occurrences for NOM_PP[6], 94 for PART, and 100 for PASS. SC requires a comparison between the frequency of the lemma in the construction and the frequency of the same lemma in the corpus, hence general frequencies for the selected lexical item were also retrieved with simple searches on *SketchEngine.* CC was conducted on the remaining construction MOD, to explore the attraction of modal + main verb in the construction. For the CC analysis, a frequency list of all the pairings of the two words in the two slots with their frequency of occurrence was retrieved. The list was manually cleaned and resulted in 1915 individual pairings of modal+ verb, and 494 significantly associated covarying collexemes. Appendix 1 reports the significant results for the two analyses.

Comparative analysis

Accessibility of CorIELLS: a comparison with the New General Service List

Having found the most significantly attracted lexical items to the 4 constructions, we explore the degree of lexical specialization in the collexemes (to answer RQ 1). To do this, we check each collexeme against the English core vocabulary in the *New General Service List* (Brezina and Gablasova 2015) (henceforth: NGSL). The NGSL is a list of ~2500 words obtained by comparing overlaps across four corpora (*Lancaster-Oslo-Bergen Corpus, British National Corpus, Corpus of British English*, and *EnTenTen12*). It aims to represent the core vocabulary of contemporary English, covering more than 80% of the text in the source corpora.

For this reason, we approximate absence from the NGSL as an indication of lexical specialization. Although the literature has acknowledged that the distinction between general and specialist lexicon is not straightforward (Bonin *et al.* 2010), this working distinction between highly accessible and less common lexicon is sufficient for the purpose at hand. Table 2 outlines the composition of the dataset and the results of the analysis in percentages.

Constructions	**Collexemes**	**%presence**	**%absence**
MOD	94	75.5%	24.5%
NOM_PP	50	66%	34%
PART	94	79.8%	20.2%
PASS	100	77%	23%

Table 2. Size of the dataset and percentages of presence/absence from NGSL

Results show that for 3 out of 4 constructions, between 20 and 25% of collexemes are not present in the NGSL, with PART being the most accessible (20.2% of specialized collexemes). NOM_PP instead shows a significantly higher percentage (34%) of specialized lexical items.[7]

The picture painted by these preliminary results is of a 'blended' genre: constructions are highly associated with accessible lexical items and highly specialized collexemes alike. This finding supports our hypothesis that decades of research on the accessibility of legal language has made LLL an autonomous and independent textual type, with idiosyncratic elements and lexico-grammatical features.

The nature of CorIELLS: a comparison with legal jargon and written prose

So far, the linguistic features of LLL have been discussed as they are found in CorIELLS. However, it is essential to also contrast LLL to other textual types to foreground how this genre is (or isn't) different from its 'parent' genre, specialized legal language.

Hence, we carry out a comparative analysis which contrasts LLL with two other genres: specialist legal jargon and general domain written prose. To do so, we use two specialized subcorpora: for legal language, an ad-hoc subcorpus of the *EurLEX* (Baisa et al., 2016) (Baisa *et al.* 2016) corpus including legislative documents in English ranging from the 90s to 2015 (henceforth: EUR)[8]; for general written language, the *BNC* imaginative subcorpus ((BNC Consortium 2007); henceforth: BNC_imag). Narrative was chosen as a proxy for non-specialist written prose since fiction is inherently aimed at large and varied audiences, and hence the use of highly specialised registers is rare. At the same time, fiction is a written genre – akin in this sense to legal language and LLL alike. All the corpora were accessed via the *SketchEngine* web interface.

Data for the comparative analysis are the above mentioned statistically associated collexemes (see Simple collexeme (SC) and Covarying collexeme (CC) analysis and Appendix 1). Frequencies of the same lexico-syntactical patterns were retrieved from both BNC_imag and EUR using CQL queries. The boxplots in Figure 1 visually represent (log transformed) frequency distributions of collexemes in constructions across the three corpora.[9] As can be seen, NOM_PP and PASS (as abstract grammatical patterns) are used very similarly in LLL (in green) and specialized legal jargon (in blue). PART and MOD instead display idiosyncratic patterns of behaviour in each corpus.

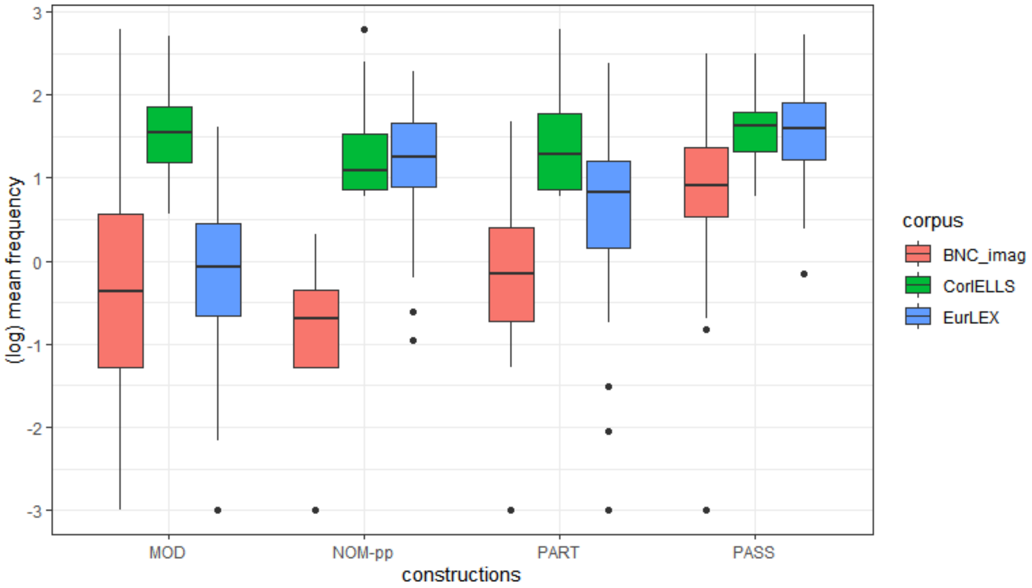

Figure 1. Frequency distributions for each construction

To test for the statistical significance of those trends, linear mixed effect modelling was used (Kuznetsova *et al.* 2017). Data were log-transformed to fit into a log-normal distribution. The model's predictors include corpora and construction type, in interaction types (i.e., in R syntax *corpus * construction*). The random intercept structure of the model includes the variable of collexeme – i.e., the different lexical items tested (in R syntax *(1| collexeme)*). Model selection was performed via Likelihood Ratio Test (Singmann *et al.* 2020). Contrasts for the variable of construction type were sum coded, i.e. "each coefficient compares the corresponding level of the factor to the average of the other levels" (Fox and Weisberg 2011: 130). In this way, the reference level for the variable (the intercept) is the overall average value for the predictor. Since we do not have a theoretically driven motivation to compare all constructions to one specific construction, this choice is the most methodologically sound. For the variable of corpus, instead, EUR was chosen as the reference level, as we are interested in analysing LLL as compared to specialized legal jargon. Therefore, all levels of all variables are statistically contrasted to average frequency mean of the four constructions in the EUR corpus.
Findings (see Table 3[10]) confirm that CorIELLS displays general overall frequency

patterns which are significantly different from legal language and written language alike, and three constructions show with idiosyncratic behaviour with respect to legal jargon.

corpus	**Predictors**	**Estimates**	**CI**	**p**
EUR	(Intercept)	0.81^{***}	0.72 – 0.91	**<0.001**
	MOD	-0.97^{***}	-1.13 – -0.81	**<0.001**
	NOM-pp	0.41^{***}	0.21 – 0.61	**<0.001**
	PART	-0.15^{*}	-0.30 – 0.00	0.05
	PASS	0.71^{***}	0.56 – 0.86	**<0.001**
BNC imag	MOD	0.76^{***}	0.55 – 0.96	**<0.001**
	NOM-pp	-1.24^{***}	-1.50 – -0.98	**<0.001**
	PART	0.06	-0.14 – 0.26	0.6
	PASS	0.43^{***}	0.22 – 0.63	**<0.001**
	MOD	1.08^{***}	0.88 – 1.28	**<0.001**
Coriells	NOM-pp	-0.57^{***}	-0.83 – -0.31	**<0.001**
	PART	0.12	-0.08 – 0.33	0.2
	PASS	-0.64^{***}	-0.84 – -0.43	**<0.001**
Comparison EUR-BNC_imag and EUR-Coriells	BNC_imag	-1.09^{***}	-1.21 – -0.96	**<0.001**
	CorIELLS	0.62^{***}	0.49 – 0.74	**<0.001**
Marginal R2 / Conditional R2: 0.489 / 0.561				

Table 3. Results of the statistical model

Particularly, in LLL modal verbs constructions (MOD) are used significantly more than in legal language (as shown by the absence of a negative sign in the estimates column), while nominalisations (NOM_PP) and passive constructions (PASS) are used significantly less. Since both NOM_PP and PASS are highly characteristic of specialised legal language, the result confirms that there are structural differences between the grammar of legal language and the grammar of language with legal content directed at a wider audience.

These findings align with our hypotheses: LLL exhibits lexico-grammatical features which are not totally ascribable to specialist legal jargon. Subcategorization preferences for the sample of constructions considered here point to a 'blended' genre, a result which is comparable with findings on Italian using the same procedure outlined in Busso (2022).

Readability of CorIELLS: is LLL more readable?

The analysis carried out on lexico-grammatical properties of LLL has provided preliminary evidence for our hypothesis of LLL as an independent and 'blended' genre between specialist legal jargon and general written prose.

We further tested this hypothesis by conducting an exploratory analysis of the *readability* of LLL with respect to the other 2 genres (specialist legal jargon and general domain written prose). Readability is here defined – following the literature – as "how easily written materials can be read and understood" (Richards and Schmidt 2013). Therefore, our definition of readability relates to text comprehension rather than processing (e.g., Kate *et al.* 2010).

To investigate text comprehension we employ readability metrics, which are widely used in the scientific literature (and beyond) to assess the reading ease/difficulty of a document. Readability measures are a useful tool, although their theoretical foundations are considered to be weak (Davison and Kantor 1982). Generally, these metrics rely on superficial text-based features such as number of words per sentence, or number of characters or syllables per word – as a proxy of respectively syntactic and lexical complexity. While both important components of readability, sentence and word length are by no means exhaustive measures of readability, which comprises several other features such as cohesion, lexical sophistication, and discourse structures (Snow 2002; Crossley *et al.* 2008).

However, a number of studies report strong correlations with text comprehension criteria (Chall and Dale 1995), and have been adopted vastly in academia and beyond. Such formulas are manifold, with well over 200 different readability scores developed since the 1920s (DuBay 2004).Particularly, different fields in linguistics have variously applied a multitude of readability formulas: L2 learning (Crossley *et al.* 2011; Xia *et al.* 2019), NLP (François and Miltsakaki 2012; Crossley *et al.* 2019; Smeuninx *et al.* 2020), psycholinguistics (Dębowski *et al.* 2015; Howcroft and Demberg 2017), language teaching (Carrell 1987; Zalmout *et al.* 2016), etc. Given that readability scores have been proven useful in research despite being far from perfect measures (Conklin *et al.* 2019), we here employ classic readability scores that will be compared to native speakers' judgments to compare text-based measures of comprehension with data collected from actual speakers.

Three readability indexes were chosen: the *Flesch-Kincaid formula* (henceforth: FK, (Flesch 1979), the *Automated Reading Index* (henceforth: ARI, (Senter and Smith 1967)), and the *Coleman-Liau Index* (henceforth: Col, (Coleman and Liau 1975)). The reason for using these particular scores is their cross-comparability, as they all employ a numerical scale based on the American school system: the higher the value, the more years of education are allegedly required to understand a given text (Figure 2).

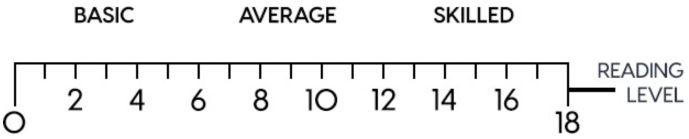

Figure 2. Grade levels used for the 3 readability measures (adapted from readable.com)

For this part of the analysis, a random sample of 20 concordances instantiating each construction was selected (total: 80 concordance lines per corpus, 240 in total). The concordances were chosen using the GDEx function on SketchEngine[11], and manually refined to have concordances of comparable length (between 100 and 200 characters, mean length 156.9 characters). Moreover, the concordances often instantiate more than one construction at a time (see Table 4). This is inevitable when working with chunks of text and not with single sentences or clauses. However, data selection was careful to include concordances with an overwhelming majority of occurrences of one construction as instances of that construction.

Table 4 reports examples of concordance lines extracted from each corpus.

The three abovementioned readability measures (FK, ARI, and COL) were calculated for the whole dataset (Rinker 2020).

	BNC_imag	**CorIELLS**	**EUR**
MOD	I thought Mr. Braden should be reminded that there were ladies present, but instead I said, "I don't know if the ladies enjoy this kind of talk very much."	If there is any inconsistency between this Part A and any other Part of these Terms and Conditions the provisions of that Part shall prevail.	For those purposes, the certification body may accompany the paying agency when it carries out secondary level on-the-spot checks.
NOM_pp	He'd appeared confident of meeting his commitments with the tourists at lunchtime at The Randolph, and then again during the afternoon.	Directive on the strengthening of certain aspects of the presumption of innocence and of the right to be present at the trial in criminal proceedings.	Agreement on the Accession of the Republic of Austria to the Convention implementing the Schengen Agreement of 14 June 1985.
PART	It was now a warm, clear night with just a soft breeze rustling the ropes and canvas of the small boats berthed in the marina far below.	We will organise a day for installation which is convenient for both of us and we will send you a letter confirming the date of your engineer appointment.	Thus, a horizontal law implementing a European directive would take precedence over conflicting provisions contained in national legislation.
PASS	Doreen was the type of girl who always sounded as though her nasal passages were obstructed or her throat sore.	Details of your normally available download speed and minimum download speed will have been provided to you at point of sale.	Any moneys recovered from loan losses for which payment has been made under guarantees called shall be credited to the Trust Account.

Table 4. Concordances examples for all four constructions

The three sets of readability scores were averaged to obtain a "meta-measure". Figure 3 plots the distribution of the averaged readability scores per corpus. As can be easily seen, the intermediate 'mixed' character of LLL seems to hold also in terms of readability, although the median value for both CorIELLS and EUR is very high (respectively, 13.9 and 14.9) with respect to BNC_imag (10.5).

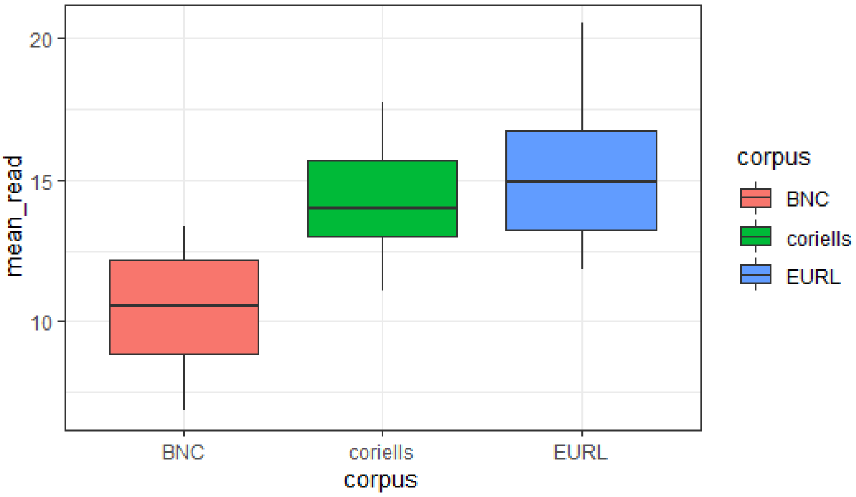

Figure 3. Boxplot of (averaged) readability per each corpus

Statistical significance was again assessed with linear mixed effect modelling, using LRT for model selection. The final model includes the variable of corpus (sum coded) as a predictor and the variable of concordance as the intercept random factor.

Table 5 outlines estimates for fixed effects and Figure 4 plots such estimates.

Predictors	Estimates	CI	standardized CI	p
(Intercept)	13.35***	12.81 – 13.90	-0.16 – 0.21	**<0.001**
BNC_imag	-2.79***	-3.55 – -2.04	-1.22 – -0.70	**<0.001**
CorIELLS	0.88*	0.08 – 1.68	0.03 – 0.58	**<0.05**
EUR	1.91***	1.16 – 2.67	0.40 – 0.92	**<0.001**

Table 5. Results for the model

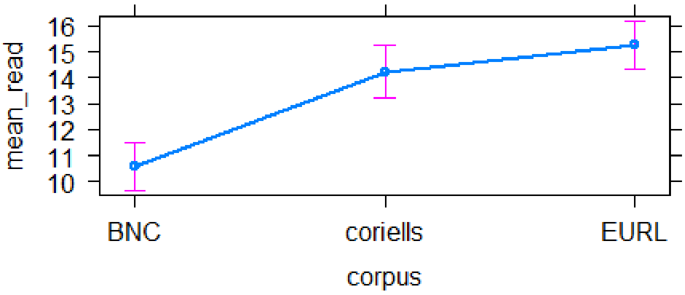

Figure 4. Estimates plot from the model

The statistical model confirms the trend found in the raw data: LLL shows a readability 0.88 grades higher than the overall mean, while legal jargon requires almost 2 grades more to be understood, while general written prose almost 3 grades less.

Difficulty of CorIELLS: can speakers understand LLL?

Readability metrics are a useful proxy for text comprehension and – to some extent – lexical and syntactical complexity. However, as we have seen in the previous paragraph, they have several limitations. Therefore, we compare the text-based analysis of readability with data collected from native speakers of English, which were presented a survey using a selection of the same concordance lines. Specifically, a random sample of concordance lines was extracted from the dataset used for the readability analysis. A total of 80 stimuli (20 BNC_imag + 20 EUR + 40 CorIELLS, 10 per subcorpus) was selected.[12] The stimuli for the survey are sentences between 100 and 200 characters long (normalised per length, mean= 156.8, st. dev= 21.6) that include one (or more) of the grammatical constructions analysed. Similarly to Table 4 above, examples from all 3 corpora and for all 4 constructions are reported in Table 6.

The survey was presented to 50 native British English speakers using the *SurveyMonkey Audience* platform[13]. Due to non-completion of task, data from 7 subjects had to be excluded from all following analyses, leaving a total of 43 participants (24 F, 19 M, median age group:31-45).

Before the survey, informed consent and a brief sociolinguistic questionnaire asking for information on gender, age, and education level were presented.[14] Stimuli were preceded by the following instructions:

> "How difficult to understand is the following sentence(s)? Use the slider to indicate how complex and difficult to understand you find the following texts, and list from 1 to 4 how well you think you understood its meaning; keep in mind that you will be show excerpts of longer texts."

Participants were then presented with the stimuli in random order, and ratings were formulated against a graded scale from 1 to 100.

Mixed models were chosen once again here as a statistical technique to control for the random effect of participants and stimuli selection. During model selection via LRT, the effect of the different constructions was found to be non-significant (p =.48), hence the final model only includes the predictor of corpus (p <.0001), with CorIELLS set as the intercept level. Ratings are log-transformed to fit a normal distribution. Random structure includes intercepts for both participant and stimuli (in R syntax, *(1|participant) + (1|stimuli)*).

Not surprisingly, results are in line with all previous analyses: general-domain prose appears to be significantly less difficult than LLL (-0.4), and legal jargon significantly more difficult (0.15) (see Table 7). This is somewhat an expected result, but still important in itself: native speakers' intuition and text comprehension confirms the corpus-based analyses described in the previous paragraphs.

Bridging the gap between readability and speakers' judgments

To compare results from the two analyses, readability scores and difficulty ratings were normalised on a common scale from 1 to 10. Figure 5 plots the (aggregated) normalised results.

The raw data from both experiments (text-based readability and human judgments) show very similar trends. However, to see if the descriptive trend can be generalised, a two- way ANOVA was carried out, with experimental condition (i.e., survey or

	BNC_imag	**CorIELLS**	**EUR**
MOD	She couldn't take their mother's place, of course, but for Liz's sake she must try to do everything she possibly could for the little girls.	The Content you submit must not include third-party intellectual property such as copyrighted material unless you have permission from that party or are otherwise legally entitled to do so.	The authorities of the Côte d'Ivoire shall communicate, before the entry into force of the Agreement, all information concerning the bank account to be used for the payment of the fees.
NOM_ pp	Even so, it was plain from the mixture of resentment and hostility on his face that her words had wounded him.	Provisions on the application and development of the Schengen acquis, relating to the abolition of checks at internal borders and movement of persons.	The Commission has examined France's application for the approval of amendments to the specification of the protected designation of origin 'Olives noires de Nyons'.
PART	The three boys sat under heavy guard in a glow-globe-lit room hung with a tapestry depicting the march across the wastes three centuries earlier.	Our 5G services may be affected by the number of people using the 5G service, maintenance and upgrades, faults from other networks, the weather, other environmental factors or degradation.	The data are based upon the "special trade" system, according to which, external trade comprises goods crossing the customs border of the country.
PASS	In the first crime, he had been robbed of something on which he had set his heart, in the second he was robbed of his life.	Any claim dispute or matter arising under or in connection with this User Agreement shall be governed and construed in all respects by the laws of England and Wales.	Bee-keeping products can only be sold as organic products if the general conditions on feeding, care and housing have been observed for at least one year.

Table 6. Example stimuli for all constructions in the 3 corpora

Predictors	Estimates	CI	standardized CI	p
CorIELLS	1.72	1.67 – 1.77	1.67 – 1.77	**<0.001**
BNC_imag	-0.39	-0.48 – -0.31	-0.48 – -0.31	**<0.001**
EUR	0.15	0.07 – 0.24	0.07 – 0.24	**<0.001**
Marginal R2 / Conditional R2 0.300 / 0.512				

Table 7. Results of the model

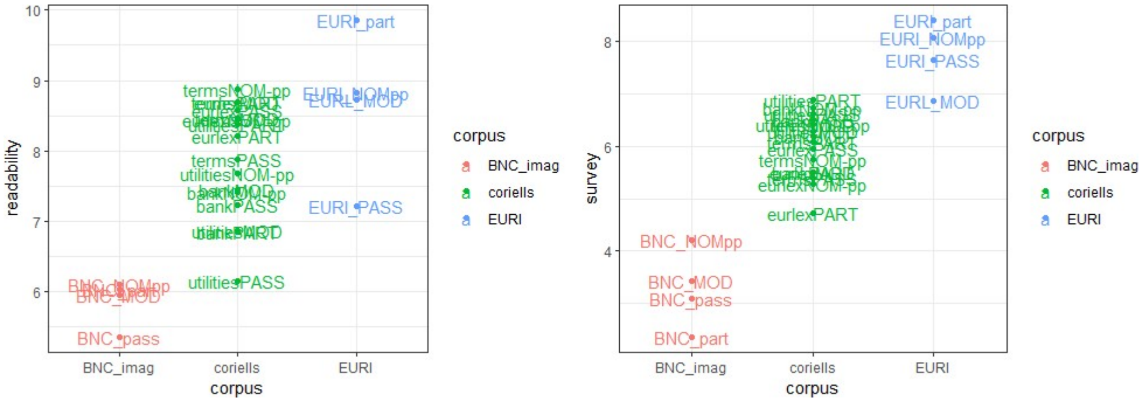

Figure 5. Normalised aggregated readability scores and difficulty ratings from the survey

readability) in interaction terms with the variable of corpus. ANOVA was chosen as a statistical method to estimate how the mean of readability and human judgments' scores is affected by the levels of the two independent categorical variables "experimental condition" and "construction".

Main effects report a significant difference between survey and readability data (F value =74.5, p-value= <.0001) and across corpora (F value=53, p-value= <.0001). A marginally significant effect is also found in the interaction between the two variables (F value =2.7, p-value= .07). Hence, readability measures seem to underestimate the accessibility of texts, as the higher estimates indicate (see Figure 5).

A *post-hoc* Tukey HSD test reveals that pairwise comparisons of corpora across conditions reach statistical significance for BNC_imag and CorIELLS (Table 7). In other words, readability scores are significantly higher for both written prose and LLL (Figure 6, Table 8), but no difference is found in the assessment of legal jargon. Here, we hypothesize that the higher accuracy of reading metrics in evaluating legal language with respect to LLL and general domain prose could lie in the 'tuning' of the metrics themselves. In fact, readability scores have been traditionally employed to analyse the accessibility of highly specialist genres (Formisano 2015).

Conclusions

The present paper has presented some preliminary quantitative analyses on English legal-lay language (LLL) using an ad-hoc compiled specialised corpus, CorIELLS. Several types of analysis were carried out on a sample of 4 lexico-grammatical constructions (Goldberg 2019): nominalisations heading prepositional phrase attachments, modal verb constructions, participial reduced relative constructions, and passive constructions.

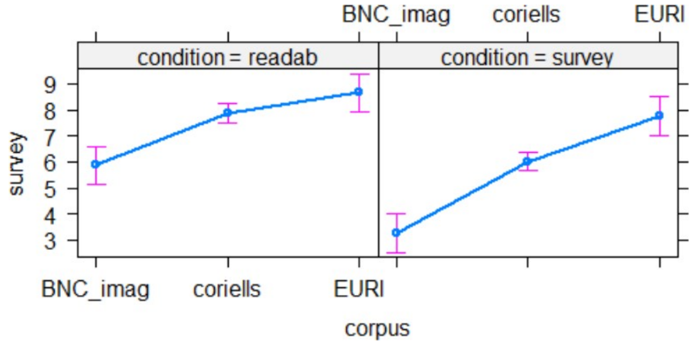

Figure 6. Effects of the ANOVA

Corpus (comparison survey*readability)	Difference	lower	upper	p-value
BNC_imag	-0.6	-0.85	-0.35	**<.0001**
CorIELLS	-0.27	-0.39	-0.14	**<.0001**
EUR	-0.11	-0.36	0.14	0.8
Marginal R2 / Conditional R2 0.300 / 0.512				

Table 8. Relevant pairwise comparisons of the post-hoc Tukey HSD test

A first exploratory part of the study set out to examine specifically the lexico-grammatical features of legal-lay English. The subcategorization preferences of the selected constructions were investigated using simple and covarying collexeme analysis (Stefanowitsch 2013). Collexemes that were found to be significantly associated with each construction were then checked against the NGSL (Brezina and Gablasova 2015) to determine the degree of specialisation of LLL. Findings indicate that NOM_pp is the construction with the most specialist lexicon out of the four constructions (34% of terms do not present in the core vocabulary).

The second part of the study aims to compare LLL to specialist legal jargon and general-domain written prose. Specifically, the frequency of statistically associated collexemes found in CorIELLS was compared with the frequency of the same structures in two other specialised corpora: the BNC imaginative subcorpus (BNC, 2007), and the English version of *EurLEX* (Baisa *et al.* 2016). Results support our hypothesis that LLL displays linguistic features quantitatively different from the other two genres. Findings point to LLL being a 'blended' genre, similarly to what was found for Italian (Busso 2022) A similar result is obtained by analysing readability scores of a sample of concordances. Interestingly, the pattern holds true also in native speakers' judgments of the same set of concordances. Even though the pattern is the same, the statistical comparison of survey responses and readability scores indicates that speakers consider legal-lay language more accessible than text-based metrics seem to suggest.

To conclude, the present study has presented the first quantitative in-depth exploration of legal-lay language, taking both a corpus-driven and an experimental perspective. Specifically, the investigation of lexico-grammatical characteristics of LLL suggests that it possesses idiosyncratic charactersistics that differentiate it from specialist legal language and general-domain written language: idiosyncratic lexical choices, and intermediate readability and comprehensibility. A comparison of text-based readability and survey data also suggests that readability metrics might underestimate the readers' ability to understand texts. However, further research in this direction is needed to confirm this preliminary finding.

Notes

[1] Different constructions differ for *schematicity*, and are distributed on a gradient cline ranging from lexical items to abstract argument structure patterns:

a. Lexical level: Word e.g. avocado, anaconda, and

b. Complex word e.g. daredevil, shoo-in

c. Complex word (partially filled) e.g. [N-s] (for plurals)

d. Phrase level: Idiom (filled) e.g. give the devil his dues, going great guns

e. Clause level: Idiom (partially filled) e.g. jog <someone's> memory

f. Covariational Conditional [The Xer the Yer] e.g. the more you think about it, the less you understand

g. Discourse level: Ditransitive (double object) [Subj V Obj1 Obj2] e.g. he gave her a fish taco, Passive [Subj aux VPPP (PPby)] e.g. the armadillo was hit by a car

[2] The corpus is freely available online on the Forensic Linguistic Databank (Petyko *et al.* 2022) https://fold.aston.ac.uk/

[3] These documents are originally drafted in English and later adapted by specialised translators and legal experts in each language of the European Union, as prescribed in EU style guides (Inter institutional Style Guide, 2015:54-62)

[4] All following examples are taken from *CorIELLS*.

[5] The CQL searches for the 4 constructions are as follows:

MODAL: [tag="MD"] [] {0,1} [tag="V.*" & tag!="VVN"]

NOMINALIZATIONS PP CHAIN: [tag= "N.*"] [tag= "IN/that|IN" & word!="and"] []?[tag="N.*"] [tag!= "SENT|SYM"] {0,2} [tag= "IN/that|IN"& word!= "and"] [ ]? [tag="N.*"] [tag!= "SENT|SYM"] {0,2} [tag="IN/that|IN" & word!="and"]

PARTICIPIAL: [tag= "NN.*"] [tag="VHG|VBG|VVG"] [tag= "DT"]

PASSIVE: [tag= "MD"]? [tag= "VB.*|VH.*"] [word= "been|being"]? [tag="VVN.*"]

[6] The general CQL simply retrieves nouns. Deverbal nouns were manually selected from the general frequency list.

[7] Interpretation of these findings was done bearing in mind that due to sparsity of linguistic data, it is inevitable for core vocabulary to cover a high percentage of the lexicon (Zipf 1949).

[8] Unfortunately, there is no easy way of knowing which of the documents in the EurLEX corpus were initially drafted in English and which one were translated from another official language.

[9] Boxplots represent data range in quartiles. The black line that divides the box into two parts is the median value (middle quartile), which marks the "mid-point" of the data. Half the frequency values are greater than or equal to this value and half are less. The first and fourth quartile are represented as the "whiskers" of the plot, while the second and third quartiles by the box.

[10] Adjusted R2 values are automatically retrieved with the package *sjplot* (Lüdecke 2021). R2 values describe the amount of variance in the data that is explained by the model. In this case, more than 56% of the variance is explained by the predictors in this model.

[11] GDEX stands for Good Dictionary Examples, a function the user can select in KWIC searches in Sketch Engine. GDEX automatically identifies sentences that are illustrative and representative of the query.

[12] The survey presented a subset of all concordance lines to avoid fatigue in participants and promote completion of the task.

[13] Available online at http://www.surveymonkey.com

[14] Sociolinguistic variables will be explored in further research, but for the purposes of this study, we will only consider corpus and construction as independent variables of interest.

References

Adler, M. (2012). *The Plain Language Movement.* Oxford: Oxford University Press.

Aher, M. (2013). *Modals in Legal Language.* University of Osnabrück.

Almut, K. (2010). Building small, specialised corpora. In O. A. and M. M, Eds., *The Routledge handbook of corpus linguistics.* London: Routledge, 66–79.

Baisa, V., Michelfeit, J., Medveď, M. and Jakubíček, M. (2016). European Union language resources in sketch engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2799–2803.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In M. Lino, Ed., *Proceedings of the Fourth Language Resources and Evaluation Conference, LREC2004 Lisbon*, 1313–1316.

Bhatia, V. (1983). Simplification v. easification— the case of legal texts. *Applied Linguistics*, 4(1), 42–54.

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

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4), 243–257.

BNC Consortium, (2007). *The British National Corpus.* XML Edition, Oxford Text Archive.

Bonin, F., Dell'Orletta, F., Venturi, G. and Montemagni, S. (2010). Singling out legal knowledge from world knowledge. an nlp–based approach". In *Proceedings of the 4th Workshop on Legal Ontologies and Artificial Intelligence Techniques*, 39–50.

Brezina, V. and Gablasova, D. (2015). Is there a core general vocabulary? introducing the new general service list. *Applied Linguistics*, 36(1), 1–22.

Bulatović, V. (2013). Legal language: The passive voice myth. *ESP Today*, 1(1), 93–112.

Busso, L. (2022). Lexicon and grammar in legal-lay language: a quantitative corpus study on Italian. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, LI(1), 5–33.

Busso, L. (forthcoming). CorIELLS: a specialised bilingual corpus of English and Italian legal-lay language. In *Presentation at the workshop Forensic linguistics between scientific research and legal practice, LIV International Conference of the Italian Linguistics Society.*

Bybee, J. (2015). *Language change.* Cambridge: Cambridge University Press.

Cappelle, B. and Depraetere, I. (2016). Modal meaning in construction grammar. *Constructions and Frames*, 8(1), 1–6.

Carrell, P. (1987). Readability in esl. *Reading in a Foreign Language*, 4, 21–40.

Chall, J. and Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula.* Cambridge, MA: Brookline Books.

Charrow, R. and Charrow, V. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia law review*, 79, 1306–1347.

Chovanec, J. (2013). Grammar in the law. In C. Chappelle, Ed., *The Encyclopedia of Applied Linguistics.* Oxford: Blackwell, 1– 8.

Coleman, M. and Liau, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60, 283–284.

Conklin, K., Hyde, R. and Parente, F. (2019). Assessing plain and intelligible language in the Consumer Rights Act: a role for reading scores? *Legal Studies*, 39(3), 378–397.

Coulthard, M. and Johnson, A. (2007). *An introduction to forensic linguistics: Language in evidence.* London: Routledge.

Crossley, S., Allen, D. and McNamara, D. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, 23(1), 84–101.

Crossley, S., Greenfield, J. and McNamara, D. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3), 475–493.

Crossley, S., Skalicky, S. and Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4), 541–561.

Davison, A. and Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2), 187–209.

DuBay, W. (2004). The principles of readability, impact information. http://www.nald.ca/library/research/readab/readab.pdf.

Dębowski, , Broda, B., Nitoń, B. and Charzyńska, E. (2015). Jasnopis–a program to compute readability of texts in Polish based on psycholinguistic research. In B. Sharp, W. Lubaszewski and R. Delmonte, Eds., *Natural Language Processing and Cognitive Science.* Venice: Libreria Editrice Cafoscarina, 51–61 ,.

T. Fanego and P. Rodríguez-Puente, Eds. (2019). *Corpus-based research on variation in English legal discourse, 91.* Amsterdam: John Benjamins.

Flach, S. (2018). Collostructions: An R implementation for the family of collostructional methods. R package version 0.1.2.

Flesch, R. (1979). *How to write in plain English: A book for lawyers and consumers.* New York: Harper.

Formisano, V. (2015). XVIII. new methodological approaches to readability corpus linguistics and dictionaries of science and technology. In G. Dotoli, C. Saggiomo, R. Spiezia and C. Boccuzzi, Eds., *La lisibilité du dictionnaire.* Paris: Hermann, 273–289.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression.* New York: Sage Publications.

Frade, C. (2007). Power dynamics and legal English. *World Englishes*, 26(1), 48–61.

Frankenreiter, J. and Livermore, M. (2020). Computational methods in legal analysis. *Annual Review of Law and Social Science*, 16, 39–57.

François, T. and Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, 49–57.

Givón, T. (1993). *English grammar: A function-based introduction*, volume 2. Amsterdam: John Benjamins.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Goldberg, A. (2019). *Explain Me This.* Princeton: Princeton University Press.

Gotti, M. (2014). Linguistic insights into legislative drafting. *The Theory and Practice of Legislation*, 2(2), 123–143.

Goźdź Roszkowski, S. and Pontrandolfo, G. (2015). Legal phraseology today: corpus-based applications across legal languages and genres. *Fachsprache: Internationale Zeitschrift für Fachsprachenforschung- didaktik und Terminologie*, 37(3), 130–139.

S. Goźdź-Roszkowski and G. Pontrandolfo, Eds. (2017). *Phraseology in Legal and Institutional Settings: A Corpus-based Interdisciplinary Perspective.* London: Routledge.

Gries, S. (2013). Data in construction grammar. In T. Hoffmann and G. Trousdale, Eds., *The Oxford handbook of Construction Grammar*. Oxford: Oxford University Press, 93–108.

Groom, N. (2019). Construction grammar and the corpus-based analysis of discourses: The case of the WAY IN WHICH construction. *International Journal of Corpus Linguistics*, 24(3), 291–323.

Haapio, H. (2011). Contract clarity through visualization: Preliminary observations and experiments. In *Proceedings of the 15th International Conference on Visualization*: IEEE Computer society.

Haigh, R. (2013). *Legal English*. London: Routledge.

Hamann, H., Vogel, F. and Gauer, I. (2016). Computer assisted legal linguistics. In F. Bex and S. Villata, Eds., *Legal Knowledge and Information Systems: JURIX 2016: The Twenty-Ninth Annual Conference*, 195–198, Amsterdam: IOS Press.

Hilpert, M. (2013). *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University Press.

Howcroft, D. and Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, 958–968: Long Papers.

Jaeggli, O. (1986). *Passive*. Linguistic inquiry.

Janigová, S. (2008). *Syntax of-ing Forms in Legal English*, volume 439. Bristol: Peter Lang.

Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R. and Mooney, R. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 546–554: Association for Computational Linguistics.

Kilgarriff, A., Rychlý, P., Smrz, P. and Tugwell, D. (2004). *Itri-04-08 The Sketch Engine*. Information Technology.

Kuznetsova, A., Brockhoff, P. and Christensen, R. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26.

Langacker, R. (2013). Modals: Striving for control. In J. Marín-Arrese, M. Carretero, J. Hita and J. Auwera, Eds., *English modality: Core, periphery and evidentiality*. Amsterdam: De Gruyter Mouton, 3– 55.

Lintao, R. and Madrunio, M. (2015). Analyzing the lexical structures of a Philippine consumer-finance contract. *Journal of Teaching English for Specific and Academic Purposes*, 2(3), 359–370.

Lüdecke, D. (2021). sjPlot: Data visualization for statistics in social science. R package version 2.8.7. https://CRAN.R-project.org/package=sjPlot.

Petyko, M., Atkins, S., Busso, L. and Grant, T. (2022). The forensic linguistic databank. *Journal of Language and Law*.

Pollard, C. and Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Richards, J. and Schmidt, R. (2013). *Longman dictionary of language teaching and applied linguistics*. London: Routledge.

Rinker, T. (2020). qdap: Quantitative discourse analysis package 2.4.2. https://github.com/trinker/qdap.

Romm, T. (2018). Facebook didn't read the terms and conditions for the app behind Cambridge Analytica. *The Washington Post.* https://www.washingtonpost.com/news/the-switch/wp/2018/04/26/ facebook-didnt-read-the-terms-and-conditions-for-the-app-behind-cambridge-analytica/ accessed on 31/3/2022.

Saussure, F. (1916). *Course in general linguistics.* London: Duckworth.

Schiess, W. (2007). The art of consumer drafting. *Scribes Journal of Legal Writing*, 11, 1–24.

Senter, R. and Smith, E. (1967). Automated readability index. Amrl-Tr. aerospace medical research laboratories (u.s). *Wright-Patterson Air Force Base*, 1–14.

Singmann, H., Bolker, B., J., A., F., and Ben-Shachar, M. (2020). afex: Analysis of factorial experiments. R package version 0.28-0. https://CRAN.R-project.org/package=afex.

Smeuninx, N., Clerck, B. and Aerts, W. (2020). Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and NLP. *International Journal of Business Communication*, 57(1), 52–85.

C. Snow, Ed. (2002). *Reading for understanding: Toward an R & D program in reading comprehension.* Santa Monica, CA: Rand.

Stefanowitsch, A. (2013). Collostructional analysis. In T. Hoffmann and G. Trousdale, Eds., *The Oxford handbook of construction grammar.* Oxford: Oxford University Press.

Stefanowitsch, A. and Gries, S. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–43.

Stefanowitsch, A. and Gries, S. (2005). Co-varying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1), 1–43.

Tiersma, P. (1999). *Legal language.* Chicago: University of Chicago Press.

Torres–Martínez, S. (2019). Taming english modals. *English Today*, 35(2), 50–57.

Van Boom, W., Desmet, P. and Dam, M. (2016). If it's easy to read, it's easy to claim'—the effect of the readability of insurance contracts on consumer expectations and conflict behaviour. *Journal of Consumer Policy*, 39(2), 187–197.

Williams, C. (2004). Legal English and plain language: an introduction. *ESP Across Cultures*, 1, 111– 124.

Williams, C. (2010). Functional or dysfunctional? the language of business contracts in english. *Rassegna Italiana di Linguistica Applicata*, 217–227.

Williams, C. (2013). Changes in the verb phrase in legislative language in english. In B. Aarts, J. Close, G. Leech and S. Wallis, Eds., *The Verb Phrase in English: Investigating Recent Language Change with Corpora (Studies in English Language).* Cambridge: Cambridge University Press, 353–371 ,.

Xia, M., Kochmar, E. and Briscoe, T. (2019). Text readability assessment for second language learners. arXiv preprint arXiv:1906.07580.

Yokoyama, S., Okamoto, H., Miyamoto, T., Yoshimoto, K., Kim, J., Iwata, K. and Kawashima, R. (2006). Cortical activation in the processing of passive sentences in L1 and L2: An fMRI study. *Neuroimage*, 30(2), 570–579.

Yunus, K. and Ab Rashid, R. (2016). Colligations of prepositions: Essential properties of legal phraseology. *International Journal of Applied Linguistics and English Literature*, 5(6), 199– 208.

Zalmout, N., Saddiki, H. and Habash, N. (2016). Analysis of foreign language teaching methods: An automatic readability approach. In *Proceedings of the 3rd workshop on natural language processing techniques for educational applications*, volume NLPTEA2016, 122–130.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Boston: Addison-Wesley.

Ződi, Z. (2019). The limits of plain legal language: understanding the comprehensible style in law. *International Journal of Law in Context*, 15(3).

Appendix 1: significantly associated collexemes from SC and CC analysis

Simple Collexeme Analysis

CONSTRUCTION	COLLEXEME	CORPUS. FREQ.	OBS.	EXP.	COLL.STR. (LOGL)	SIGNIF
NOM-pp	use	3518	507	48.1	1553.591	*****
	charge	610	204	8.3	988.6324	*****
	purchase	231	67		301.9074	*****
	transfer	412	59	5.6	178.0778	*****
	refund	178	38	2.4	145.6235	*****
	display	171	32	2.3	113.7878	*****
	change	603	48	8.2	92.50596	*****
	accordance	532	34	7.3	52.88384	*****
	conclusion	112	16	1.5	48.1655	*****
	protection	334	23	4.6	38.60167	*****
	report	155	16	2.1	38.25826	*****
	processing	164	14	2.2	28.66795	*****
	access	671	29	9.2	27.75447	*****
	payment	2093	60	28.6	26.67419	*****
	application	484	22	6.6	22.63027	*****
	consideration	31	6	0.4	21.74256	*****
	relief	36	6	0.5	19.90199	*****
	impact	78	8	1.1	19.02998	****
	transmission	79	8	1.1	18.84251	****
	provision	496	20	6.8	17.21279	****
	booking	144	10	2	16.91684	****
	assistance	97	8	1.3	15.89314	****
	notice	919	28	12.6	14.30254	***
	obligation	557	20	7.6	14.15943	***
	indemnification	39	5	0.5	13.99496	***
	information	2859	64	39.1	13.58679	***
	participation	75	6	1	11.59979	***
	connection	532	18	7.3	11.40833	***
	assessment	68	5	0.9	8.93997	**
	accommodation	82	5	1.1	7.38728	**
	loss	628	17	8.6	6.52525	*
	procedure	300	10	4.1	6.15402	*
	notification	186	7	2.5	5.37681	*
	supplier	340	10	4.6	4.71098	*
	agreement	2352	45	32.1	4.65463	*

CONSTRUCTION	COLLEXEME	CORPUS. FREQ.	OBS.	EXP.	COLL.STR. (LOGL)	SIGNIF
	addition	167	6	2.3	4.24951	*
	restriction	173	6	2.4	3.98334	*
PART	make	2264	519	30.9	2088.002	*****
	require	704	277	9.6	1452.694	*****
	arise	370	211	5.1	1314.178	*****
	relate	514	231	7	1288.317	*****
	provide	2482	369	33.9	1149.601	*****
	govern	260	164	3.6	1070.565	*****
	regard	194	128	2.7	853.4201	*****
	pay	1615	260	22.1	849.2509	*****
	give	1144	211	15.6	746.8929	*****
	use	3518	326	48.1	722.0198	*****
	apply	1523	224	20.8	690.6339	*****
	take	1146	188	15.7	620.1748	*****
	send	738	157	10.1	601.9633	*****
	set	1025	169	14	558.9743	*****
	receive	1018	155	13.9	487.7473	*****
	process	324	90	4.4	396.9423	*****
	hold	384	84	5.2	326.5629	*****
	carry	353	75	4.8	286.8725	*****
	result	151	56	2.1	284.5307	*****
	enter	377	66	5.2	225.8185	*****
	remove	284	53	3.9	188.245	*****
	determine	173	42	2.4	172.5715	*****
	read	222	45	3	167.5523	*****
	request	312	51	4.3	167.3401	*****
	exclude	200	43	2.7	165.4476	*****
	grant	271	48	3.7	165.311	*****
	ask	556	63	7.6	161.7833	*****
	meet	219	42	3	151.5138	*****
	label	72	28	1	145.4475	*****
	confirm	196	38	2.7	137.9314	*****
	follow	626	60	8.6	135.4979	*****
	affect	321	44	4.4	128.9828	*****
	establish	349	45	4.8	126.5791	*****
	post	299	41	4.1	120.21	*****
	offer	319	42	4.4	119.8267	*****
	cover	307	41	4.2	118.1031	*****
	amend	313	41	4.3	116.5653	*****
	show	187	31	2.6	102.5493	*****
	depend	193	31	2.6	100.5764	*****
	share	261	33	3.6	91.55111	*****
	message	44	17	0.6	88.01939	*****
	contain	280	33	3.8	87.12635	*****
	display	283	32	3.9	81.98178	*****

CONSTRUCTION	COLLEXEME	CORPUS. FREQ.	OBS.	EXP.	COLL.STR. (LOGL)	SIGNIF
	allow	490	40	6.7	78.82668	*****
	handle	59	15	0.8	63.11487	*****
	originate	23	10	0.3	54.73058	*****
	operate	293	25	4	51.1886	*****
	ship	23	9	0.3	46.87404	*****
	remain	256	22	3.5	45.32067	*****
	include	2379	77	32.5	44.83004	*****
	act	582	32	8	42.07543	*****
	involve	140	16	1.9	41.29394	*****
	tamper	40	9	0.5	35.47752	*****
	implement	158	15	2.2	33.5704	*****
	block	116	13	1.6	33.07002	*****
	work	388	23	5.3	32.96131	*****
	fall	34	8	0.5	32.3056	*****
	belong	49	8	0.7	26.20415	*****
	go	244	16	3.3	25.54625	*****
	exploit	39	7	0.5	24.27617	*****
	comprise	20	5	0.3	20.84979	*****
	exceed	138	10	1.9	17.63473	****
	accompany	52	6	0.7	15.58929	****
	travel	52	6	0.7	15.58929	****
	seek	106	8	1.4	14.66188	***
	maintain	169	10	2.3	14.29531	***
	appear	67	6	0.9	12.79405	***
	build	48	5	0.7	12.03575	***
	host	54	5	0.7	10.96101	***
	copy	55	5	0.8	10.79606	**
	live	118	7	1.6	10.034	**
	report	120	7	1.6	9.84651	**
	begin	65	5	0.9	9.32652	**
	visit	137	7	1.9	8.4067	**
	indicate	74	5	1	8.22817	**
	enable	188	8	2.6	7.47512	**
	open	193	8	2.6	7.18361	**
	address	90	5	1.2	6.64826	**
	deal	95	5	1.3	6.22998	*
	describe	213	8	2.9	6.12416	*
	order	114	5	1.6	4.88416	*
	start	158	6	2.2	4.67928	*
	exercise	122	5	1.7	4.41124	*
PASS	entitle	243	241	3.3	2051.113	*****
	find	384	264	5.2	1798.968	*****
	deem	166	113	2.3	764.802	*****
	lose	155	103	2.1	688.8913	*****
	limit	420	117	5.7	517.032	*****

CONSTRUCTION	COLLEXEME	CORPUS. FREQ.	OBS.	EXP.	COLL.STR. (LOGL)	SIGNIF
	return	257	99	3.5	512.5376	*****
	base	265	97	3.6	490.106	*****
	bind	96	69	1.3	479.4927	*****
	consider	195	78	2.7	410.9374	*****
	terminate	365	88	5	360.5733	*****
	prohibit	126	60	1.7	342.8743	*****
	authorise	299	78	4.1	333.0243	*****
	register	278	75	3.8	325.7948	*****
	calculate	92	51	1.3	312.7697	*****
	accept	464	83	6.3	287.7518	*****
	add	211	61	2.9	274.3903	*****
	conduct	83	45	1.1	273.1023	*****
	resolve	175	54	2.4	250.9118	*****
	design	94	44	1.3	249.3809	*****
	bring	121	47	1.7	244.0819	*****
	place	173	52	2.4	238.4778	*****
	treat	92	41	1.3	227.1079	*****
	agree	1364	113	18.6	226.1232	*****
	oblige	49	33	0.7	221.9553	*****
	commit	82	37	1.1	206.1294	*****
	protect	234	52	3.2	203.775	*****
	issue	171	46	2.3	199.4166	*****
	delay	53	31	0.7	194.9057	*****
	close	250	50	3.4	184.7813	*****
	activate	60	31	0.8	183.9247	*****
	notify	392	59	5.4	183.8804	*****
	deliver	178	44	2.4	182.5272	*****
	convert	94	35	1.3	178.1086	*****
	obtain	198	44	2.7	172.3971	*****
	collect	331	51	4.5	161.295	*****
	store	151	38	2.1	159.109	*****
	record	82	31	1.1	158.891	*****
	intend	153	38	2.1	158.0131	*****
	cancel	618	64	8.4	153.6224	*****
	list	212	41	2.9	148.6182	*****
	restrict	188	39	2.6	147.0925	*****
	connect	172	37	2.4	142.3869	*****
	supply	366	49	5	141.4103	*****
	cause	233	41	3.2	140.6321	*****
	refuse	181	37	2.5	138.3981	*****
	submit	312	45	4.3	136.4086	*****
	install	140	33	1.9	133.4536	*****
	view	117	31	1.6	133.3064	*****
	identify	165	33	2.3	121.912	*****
	update	209	36	2.9	121.9004	*****

CONSTRUCTION	COLLEXEME	CORPUS. FREQ.	OBS.	EXP.	COLL.STR. (LOGL)	SIGNIF
	inform	186	34	2.5	119.2593	*****
	suspend	254	37	3.5	112.8603	*****
	choose	319	38	4.4	101.1104	*****
	fail	213	32	2.9	99.55127	*****
	tell	562	45	7.7	87.17637	*****
	exclude	200	27	2.7	78.31423	*****
	govern	260	29	3.6	73.5456	*****
	establish	349	31	4.8	65.70088	*****
	use	1396	61	19.1	59.40919	*****
	end	902	42	12.3	44.70849	*****
	act	582	30	8	36.46254	*****
	amend	313	18	4.3	24.92328	*****
	enter	377	17	5.2	17.28779	****
	do	2403	59	32.8	17.16697	****
	cover	307	15	4.2	17.01284	****
	process	324	15	4.4	15.82365	****
	post	299	14	4.1	14.99884	***
	confirm	196	11	2.7	14.80436	***
	allow	490	17	6.7	11.29948	***
	send	738	22	10.1	10.69717	**
	include	2379	51	32.5	9.12327	**
	refund	378	13	5.2	8.49806	**
	share	261	10	3.6	7.91758	**
	take	1146	27	15.7	6.85899	**
	require	704	18	9.6	5.90119	*
	display	102	5	1.4	5.6929	*
	request	311	9	4.3	4.08111	*
	display	171	6	2.3	4.07035	*
	determine	173	6	2.4	3.98334	*
	purchase	219	7	3	3.95745	*

Covarying Collexeme Analysis

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
can	find	2902	224	193	34.1	566.8055	*****
would	like	366	58	58	1.1	467.9355	*****
must	ensure	1684	117	69	10.3	187.4634	*****
will	refund	6900	109	101	39.5	155.9121	*****
shall	be	1134	4484	430	267.1	124.8021	*****
can	guarantee	2902	38	36	5.8	120.8059	*****
will	try	6900	78	71	28.3	103.7473	*****
must	comply	1684	67	38	5.9	98.70649	*****
may	require	5113	132	89	35.4	95.07449	*****
can	ask	2902	204	89	31.1	94.7817	*****
may	include	5113	179	110	48.1	94.76683	*****
should	contact	521	224	41	6.1	94.48616	*****
will	tell	6900	226	152	81.9	90.48314	*****
must	pay	1684	292	81	25.8	89.18392	*****
could	damage	215	22	13	0.2	87.79367	*****
shall	deem	1134	79	33	4.7	85.21312	*****
may	assign	5113	53	46	14.2	84.23325	*****
must	follow	1684	38	26	3.4	81.27988	*****
will	continue	6900	142	104	51.5	80.99488	*****
will	notify	6900	171	119	62	79.03737	*****
shall	govern	1134	82	32	4.9	77.67637	*****
may	charge	5113	254	130	68.2	68.44959	*****
can	change	2902	227	83	34.6	62.76964	*****
may	arise	5113	38	33	10.2	60.44768	*****
may	suspend	5113	77	53	20.7	59.10666	*****
would	compromise	366	12	9	0.2	57.96643	*****
will	apply	6900	416	225	150.8	56.03956	*****
will	send	6900	210	129	76.1	55.37221	*****
would	have	366	732	48	14.1	55.01709	*****
may	offer	5113	66	46	17.7	52.71069	*****
may	change	5113	227	111	61	50.52549	*****
will	give	6900	316	176	114.5	50.21346	*****
can	use	2902	592	156	90.2	50.12886	*****
will	be	6900	4484	1824	1625	49.46702	*****
will	let	6900	75	57	27.2	49.45472	*****
must	inform	1684	79	29	7	46.38278	*****
shall	limit	1134	72	23	4.3	45.89666	*****
may	vary	5113	66	44	17.7	45.62345	*****
should	read	521	40	13	1.1	44.89637	*****
will	treat	6900	50	41	18.1	44.31236	*****
can	cancel	2902	162	59	24.7	44.04661	*****
shall	remain	1134	83	24	4.9	43.15618	*****

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
should	check	521	52	14	1.4	42.60423	*****
would	cause	366	39	11	0.7	41.91893	*****
may	result	5113	60	40	16.1	41.46047	*****
can	prove	2902	18	15	2.7	41.26745	*****
shall	survive	1134	29	14	1.7	40.79821	*****
can	choose	2902	69	33	10.5	40.74453	*****
could	have	215	732	31	8.3	39.83457	*****
would	prefer	366	5	5	0.1	39.58395	*****
can	do	2902	192	64	29.3	39.19543	*****
will	start	6900	44	36	15.9	38.64947	*****
can	contact	2902	224	71	34.1	38.48193	*****
may	refuse	5113	71	44	19.1	38.43534	*****
can	purchase	2902	29	19	4.4	37.5146	*****
should	know	521	15	8	0.4	37.35178	*****
might	happen	199	22	7	0.2	36.88316	*****
might	need	199	261	17	2.7	35.54358	*****
shall	prevail	1134	10	8	0.6	35.42177	*****
must	keep	1684	81	26	7.2	34.86547	*****
must	sign	1684	16	11	1.4	34.46942	*****
may	differ	5113	13	13	3.5	34.20785	*****
can	access	2902	66	30	10.1	33.98812	*****
can	learn	2902	9	9	1.4	33.88424	*****
shall	preclude	1134	6	6	0.4	33.8794	*****
can	get	2902	53	26	8.1	33.42554	*****
may	request	5113	66	40	17.7	33.07173	*****
can	withdraw	2902	29	18	4.4	32.93825	*****
may	terminate	5113	113	59	30.3	32.70531	*****
shall	conduct	1134	28	12	1.7	31.52166	*****
would	expect	366	12	6	0.2	31.11062	*****
can	transfer	2902	132	46	20.1	31.10582	*****
should	note	521	6	5	0.2	30.68053	*****
can	make	2902	367	96	55.9	29.64284	*****
will	explain	6900	19	18	6.9	29.63363	*****
can	close	2902	75	31	11.4	29.63221	*****
will	need	6900	261	137	94.6	28.96456	*****
must	repay	1684	16	10	1.4	28.49675	*****
shall	cooperate	1134	13	8	0.8	28.47256	*****
may	restrict	5113	36	25	9.7	28.3646	*****
must	destroy	1684	8	7	0.7	28.1379	*****
should	exercise	521	25	8	0.7	27.28285	*****
must	meet	1684	24	12	2.1	27.22406	*****
can	recall	2902	7	7	1.1	26.35031	*****
may	monitor	5113	25	19	6.7	26.19971	*****
can	obtain	2902	31	17	4.7	25.96613	*****

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
should	direct	521	8	5	0.2	25.61219	*****
may	enable	5113	30	21	8.1	24.2416	*****
may	amend	5113	24	18	6.4	24.12743	*****
might	have	199	732	24	7.7	24.00311	*****
must	return	1684	51	17	4.5	23.93731	*****
can	refer	2902	46	21	7	23.9348	*****
will	remain	6900	83	52	30.1	23.88316	*****
would	complicate	366	3	3	0.1	23.73417	*****
should	review	521	43	9	1.2	22.66149	*****
can	promise	2902	6	6	0.9	22.58423	*****
would	prevent	366	33	7	0.6	22.3406	*****
will	confirm	6900	31	24	11.2	21.94073	*****
may	have	5113	732	253	196.6	21.89427	*****
can	see	2902	38	18	5.8	21.82263	*****
will	receive	6900	132	74	47.8	21.53734	*****
need	help	3	35	2	0	21.4965	*****
must	provide	1684	369	60	32.6	21.13179	*****
will	assume	6900	28	22	10.1	20.99686	*****
must	file	1684	18	9	1.6	20.40606	*****
will	cost	6900	10	10	3.6	20.30966	*****
will	expire	6900	10	10	3.6	20.30966	*****
may	revise	5113	11	10	3	20.23183	*****
may	suffer	5113	20	15	5.4	20.10122	*****
may	appear	5113	22	16	5.9	20.0701	*****
ought	have	3	732	3	0.1	19.56294	*****
could	lead	215	14	4	0.2	19.41475	****
would	jeopardise	366	4	3	0.1	19.27399	****
must	adhere	1684	6	5	0.5	19.0452	****
may	expose	5113	13	11	3.5	19.02817	****
shall	apply	1134	416	48	24.8	18.9655	****
must	register	1684	33	12	2.9	18.89035	****
shall	have	1134	732	73	43.6	18.57709	****
may	delegate	5113	7	7	1.9	18.41359	****
may	encounter	5113	7	7	1.9	18.41359	****
shall	entitle	1134	71	15	4.2	18.39109	****
will	endeavor	6900	9	9	3.3	18.27786	****
might	exacerbate	199	2	2	0	18.26393	****
must	satisfy	1684	9	6	0.8	18.22048	****
must	present	1684	16	8	1.4	18.13518	****
may	impose	5113	21	15	5.6	18.09136	****
shall	serve	1134	12	6	0.7	17.97708	****
may	need	5113	261	101	70.1	17.53838	****
can	email	2902	9	7	1.4	17.47563	****
may	contain	5113	37	22	9.9	17.30909	****

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
can	recover	2902	14	9	2.1	17.28422	****
can	borrow	2902	7	6	1.1	17.17271	****
can	foresee	2902	7	6	1.1	17.17271	****
would	consider	366	48	7	0.9	17.13897	****
may	modify	5113	28	18	7.5	17.11652	****
might	suffer	199	20	4	0.2	16.88158	****
might	arise	199	38	5	0.4	16.81933	****
should	report	521	25	6	0.7	16.7389	****
could	claim	215	8	3	0.1	16.47053	****
will	handle	6900	8	8	2.9	16.24625	****
may	upgrade	5113	16	12	4.3	16.077	****
would	break	366	13	4	0.2	15.95374	****
must	establish	1684	14	7	1.2	15.86518	****
must	specify	1684	14	7	1.2	15.86518	****
could	cause	215	39	5	0.4	15.8366	****
would	risk	366	2	2	0	15.81739	****
shall	deal	1134	14	6	0.8	15.73564	****
would	create	366	38	6	0.7	15.59174	****
can	control	2902	18	10	2.7	15.55951	****
shall	determine	1134	26	8	1.5	15.28585	****
will	depend	6900	65	39	23.6	15.1364	***
may	increase	5113	39	22	10.5	15.09441	***
will	post	6900	35	24	12.7	15.07687	***
can	afford	2902	4	4	0.6	15.05381	***
must	submit	1684	57	15	5	14.90264	***
must	maintain	1684	11	6	1	14.88876	***
could	disable	215	10	3	0.1	14.88306	***
must	design	1684	5	4	0.4	14.5924	***
shall	exclude	1134	21	7	1.3	14.50793	***
will	bind	6900	11	10	4	14.50672	***
should	consult	521	2	2	0.1	14.40167	***
should	fly	521	2	2	0.1	14.40167	***
should	pack	521	2	2	0.1	14.40167	***
should	speak	521	2	2	0.1	14.40167	***
can	produce	2902	8	6	1.2	14.2471	***
will	process	6900	64	38	23.2	14.134	***
may	share	5113	58	29	15.6	14.03645	***
will	calculate	6900	21	16	7.6	13.94432	***
can	visit	2902	6	5	0.9	13.74221	***
can	end	2902	138	38	21	13.72649	***
may	decide	5113	38	21	10.2	13.62585	***
could	affect	215	106	7	1.2	13.6079	***
might	break	199	13	3	0.1	13.57267	***
should	tell	521	226	17	6.2	13.52632	***

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
will	govern	6900	82	46	29.7	13.39147	***
must	review	1684	43	12	3.8	13.07601	***
can	view	2902	17	9	2.6	13.01498	***
may	assert	5113	8	7	2.1	13.00998	***
will	deduct	6900	27	19	9.8	12.97525	***
will	attempt	6900	18	14	6.5	12.96417	***
will	begin	6900	18	14	6.5	12.96417	***
will	cease	6900	18	14	6.5	12.96417	***
will	communicate	6900	18	14	6.5	12.96417	***
must	adopt	1684	13	6	1.1	12.47142	***
might	involve	199	16	3	0.2	12.23762	***
should	ensure	521	117	11	3.2	12.22537	***
could	submit	215	57	5	0.6	12.22368	***
can	enforce	2902	21	10	3.2	12.21516	***
may	access	5113	66	31	17.7	12.1992	***
will	investigate	6900	15	12	5.4	12.05844	***
shall	obligate	1134	8	4	0.5	11.97888	***
would	constitute	366	36	5	0.7	11.76027	***
could	result	215	60	5	0.7	11.75423	***
can	elect	2902	12	7	1.8	11.69988	***
might	interest	199	5	2	0.1	11.59623	***
will	credit	6900	17	13	6.2	11.45125	***
could	last	215	5	2	0.1	11.29049	***
can	chat	2902	3	3	0.5	11.28948	***
can	complain	2902	3	3	0.5	11.28948	***
shall	erase	1134	2	2	0.1	11.28648	***
would	pay	366	292	15	5.6	11.2774	***
may	update	5113	33	18	8.9	11.26035	***
can	inspect	2902	7	5	1.1	11.1034	***
will	bill	6900	19	14	6.9	11.03176	***
will	work	6900	19	14	6.9	11.03176	***
will	deem	6900	79	43	28.6	10.84535	***
shall	dispose	1134	9	4	0.5	10.82635	**
shall	indemnify	1134	9	4	0.5	10.82635	**
may	reduce	5113	31	17	8.3	10.79346	**
may	owe	5113	7	6	1.9	10.66541	**
may	undermine	5113	7	6	1.9	10.66541	**
can	accept	2902	42	15	6.4	10.64658	**
must	report	1684	25	8	2.2	10.63286	**
must	connect	1684	15	6	1.3	10.59521	**
may	consolidate	5113	4	4	1.1	10.52034	**
may	exempt	5113	4	4	1.1	10.52034	**
may	import	5113	4	4	1.1	10.52034	**
may	participate	5113	4	4	1.1	10.52034	**

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
must	tell	1684	226	35	20	10.4605	**
shall	construe	1134	5	3	0.3	10.44711	**
can	book	2902	10	6	1.5	10.44454	**
must	respect	1684	7	4	0.6	10.40532	**
can	manage	2902	13	7	2	10.38579	**
can	switch	2902	5	4	0.8	10.38011	**
may	choose	5113	69	31	18.5	10.37983	**
shall	bear	1134	15	5	0.9	10.35645	**
must	obey	1684	4	3	0.4	10.24331	**
must	proceed	1684	4	3	0.4	10.24331	**
will	acknowledge	6900	5	5	1.8	10.15252	**
will	compensate	6900	5	5	1.8	10.15252	**
will	re-credit	6900	5	5	1.8	10.15252	**
will	redirect	6900	5	5	1.8	10.15252	**
may	delay	5113	9	7	2.4	10.12834	**
may	wish	5113	9	7	2.4	10.12834	**
can	call	2902	20	9	3	9.98746	**
must	give	1684	316	45	27.9	9.9867	**
may	add	5113	53	25	14.2	9.97709	**
must	notify	1684	171	28	15.1	9.94272	**
would	mean	366	27	4	0.5	9.88797	**
must	operate	1684	21	7	1.9	9.83176	**
must	activate	1684	2	2	0.2	9.70365	**
must	seat	1684	2	2	0.2	9.70365	**
must	subscribe	1684	2	2	0.2	9.70365	**
must	tamper	1684	2	2	0.2	9.70365	**
must	travel	1684	2	2	0.2	9.70365	**
must	trust	1684	2	2	0.2	9.70365	**
could	interfere	215	7	2	0.1	9.68984	**
will	correct	6900	11	9	4	9.64517	**
will	respond	6900	11	9	4	9.64517	**
must	exceed	1684	27	8	2.4	9.52999	**
can	request	2902	66	20	10.1	9.52951	**
can	read	2902	40	14	6.1	9.5007	**
might	want	199	8	2	0.1	9.39141	**
will	pass	6900	22	15	8	9.24075	**
can	download	2902	8	5	1.2	9.22441	**
may	submit	5113	57	26	15.3	9.20654	**
would	encourage	366	5	2	0.1	9.20312	**
might	lack	199	1	1	0	9.12697	**
could	misuse	215	8	2	0.1	9.09076	**
could	relate	215	8	2	0.1	9.09076	**
can	discuss	2902	11	6	1.7	9.07676	**
could	harvest	215	1	1	0	8.97192	**

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
could	mislead	215	1	1	0	8.97192	**
could	overburden	215	1	1	0	8.97192	**
might	delay	199	9	2	0.1	8.87487	**
may	use	5113	592	191	159	8.76927	**
may	ask	5113	204	74	54.8	8.75666	**
will	convert	6900	47	27	17	8.72581	**
shall	inform	1134	79	12	4.7	8.67143	**
must	agree	1684	55	12	4.9	8.50043	**
must	disconnect	1684	13	5	1.1	8.42179	**
may	appeal	5113	6	5	1.6	8.36932	**
may	subcontract	5113	6	5	1.6	8.36932	**
may	harm	5113	12	8	3.2	8.26756	**
must	log	1684	5	3	0.4	8.19678	**
should	make	521	367	20	10	8.12266	**
will	count	6900	4	4	1.4	8.12164	**
will	honour	6900	4	4	1.4	8.12164	**
will	migrate	6900	4	4	1.4	8.12164	**
will	scan	6900	4	4	1.4	8.12164	**
will	undertake	6900	4	4	1.4	8.12164	**
would	violate	366	18	3	0.3	8.09176	**
can	avoid	2902	6	4	0.9	8.0763	**
can	phone	2902	6	4	0.9	8.0763	**
can	sit	2902	6	4	0.9	8.0763	**
may	reject	5113	10	7	2.7	8.07055	**
may	incur	5113	29	15	7.8	8.04451	**
may	record	5113	29	15	7.8	8.04451	**
will	administer	6900	10	8	3.6	8.03678	**
may	allow	5113	65	28	17.5	7.93793	**
would	hate	366	1	1	0	7.90601	**
would	outweigh	366	1	1	0	7.90601	**
would	shield	366	1	1	0	7.90601	**
would	struggle	366	1	1	0	7.90601	**
may	accompany	5113	3	3	0.8	7.88982	**
may	edit	5113	3	3	0.8	7.88982	**
can	help	2902	35	12	5.3	7.76844	**
would	affect	366	106	7	2	7.66312	**
shall	constitute	1134	36	7	2.1	7.60735	**
shall	affect	1134	106	14	6.3	7.59033	**
shall	consist	1134	3	2	0.2	7.59	**
shall	deprive	1134	3	2	0.2	7.59	**
shall	procure	1134	3	2	0.2	7.59	**
shall	relieve	1134	3	2	0.2	7.59	**
can	award	2902	2	2	0.3	7.52574	**
can	escape	2902	2	2	0.3	7.52574	**

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
can	flex	2902	2	2	0.3	7.52574	**
can	foretell	2902	2	2	0.3	7.52574	**
can	litigate	2902	2	2	0.3	7.52574	**
can	resell	2902	2	2	0.3	7.52574	**
can	sort	2902	2	2	0.3	7.52574	**
can	telephone	2902	2	2	0.3	7.52574	**
can	trace	2902	2	2	0.3	7.52574	**
can	unlock	2902	2	2	0.3	7.52574	**
might	affect	199	106	5	1.1	7.50978	**
will	renew	6900	12	9	4.3	7.47947	**
will	charge	6900	254	113	92	7.38018	**
could	expect	215	12	2	0.1	7.3647	**
could	harm	215	12	2	0.1	7.3647	**
could	subject	215	12	2	0.1	7.3647	**
might	expose	199	13	2	0.1	7.33043	**
must	call	1684	20	6	1.8	7.27446	**
should	believe	521	1	1	0	7.19896	**
should	disagree	521	1	1	0	7.19896	**
should	integrate	521	1	1	0	7.19896	**
should	intensify	521	1	1	0	7.19896	**
should	preserve	521	1	1	0	7.19896	**
should	reuse	521	1	1	0	7.19896	**
should	ring	521	1	1	0	7.19896	**
should	talk	521	1	1	0	7.19896	**
should	trigger	521	1	1	0	7.19896	**
can	earn	2902	20	8	3	7.15655	**
must	reimburse	1684	10	4	0.9	7.0603	**
would	run	366	8	2	0.2	7.05173	**
must	set	1684	75	14	6.6	7.04056	**
could	impact	215	13	2	0.1	7.03829	**
might	cause	199	39	3	0.4	7.00401	**
will	commence	6900	16	11	5.8	6.96221	**
will	reach	6900	16	11	5.8	6.96221	**
shall	waive	1134	8	3	0.5	6.95977	**
must	protect	1684	6	3	0.5	6.79405	**
should	refer	521	46	5	1.3	6.66058	**
may	violate	5113	18	10	4.8	6.57504	*
can	show	2902	42	13	6.4	6.54622	*
may	redeem	5113	11	7	3	6.49206	*
may	sell	5113	11	7	3	6.49206	*
can	buy	2902	7	4	1.1	6.48411	*
can	revoke	2902	7	4	1.1	6.48411	*
could	interpret	215	15	2	0.2	6.46549	*
must	cover	1684	11	4	1	6.28479	*

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
can	refuse	2902	71	19	10.8	6.22618	*
could	disrupt	215	2	1	0	6.22194	*
shall	sever	1134	9	3	0.5	6.21006	*
may	adjust	5113	9	6	2.4	6.19952	*
may	develop	5113	5	4	1.3	6.14134	*
may	opt	5113	5	4	1.3	6.14134	*
may	lose	5113	21	11	5.6	6.12203	*
can	go	2902	14	6	2.1	6.1042	*
will	abide	6900	3	3	1.1	6.09096	*
will	alert	6900	3	3	1.1	6.09096	*
will	defend	6900	3	3	1.1	6.09096	*
will	drop	6900	3	3	1.1	6.09096	*
will	guide	6900	3	3	1.1	6.09096	*
must	achieve	1684	3	2	0.3	6.06956	*
must	adapt	1684	3	2	0.3	6.06956	*
must	attend	1684	3	2	0.3	6.06956	*
must	declare	1684	3	2	0.3	6.06956	*
must	fulfil	1684	3	2	0.3	6.06956	*
must	suggest	1684	3	2	0.3	6.06956	*
will	come	6900	11	8	4	6.05325	*
may	invite	5113	7	5	1.9	6.02524	*
may	search	5113	7	5	1.9	6.02524	*
will	show	6900	42	23	15.2	5.96328	*
may	bring	5113	43	19	11.5	5.95787	*
may	enter	5113	32	15	8.6	5.84956	*
must	reflect	1684	7	3	0.6	5.73603	*
will	deliver	6900	21	13	7.6	5.68753	*
will	aim	6900	6	5	2.2	5.64536	*
will	state	6900	6	5	2.2	5.64536	*
shall	condemn	1134	1	1	0.1	5.64241	*
shall	confer	1134	1	1	0.1	5.64241	*
shall	excuse	1134	1	1	0.1	5.64241	*
shall	fall	1134	1	1	0.1	5.64241	*
shall	inure	1134	1	1	0.1	5.64241	*
shall	measure	1134	1	1	0.1	5.64241	*
shall	prejudice	1134	1	1	0.1	5.64241	*
shall	recredit	1134	1	1	0.1	5.64241	*
shall	twitch	1134	1	1	0.1	5.64241	*
will	collect	6900	47	25	17	5.60288	*
may	cancel	5113	162	57	43.5	5.45926	*
would	subject	366	12	2	0.2	5.39012	*
might	qualify	199	3	1	0	5.3497	*
will	write	6900	32	18	11.6	5.29098	*
can	claim	2902	8	4	1.2	5.28485	*

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
can	instruct	2902	8	4	1.2	5.28485	*
can	turn	2902	8	4	1.2	5.28485	*
could	include	215	179	6	2	5.26471	*
may	advertise	5113	2	2	0.5	5.2596	*
may	concern	5113	2	2	0.5	5.2596	*
may	contract	5113	2	2	0.5	5.2596	*
may	divert	5113	2	2	0.5	5.2596	*
may	exonerate	5113	2	2	0.5	5.2596	*
may	experience	5113	2	2	0.5	5.2596	*
may	filter	5113	2	2	0.5	5.2596	*
may	interrupt	5113	2	2	0.5	5.2596	*
may	persist	5113	2	2	0.5	5.2596	*
may	pool	5113	2	2	0.5	5.2596	*
may	recoup	5113	2	2	0.5	5.2596	*
may	stipulate	5113	2	2	0.5	5.2596	*
should	discontinue	521	9	2	0.2	5.25398	*
should	wish	521	9	2	0.2	5.25398	*
can	combine	2902	5	3	0.8	5.22023	*
would	accrue	366	2	1	0	5.17214	*
would	license	366	2	1	0	5.17214	*
shall	exceed	1134	27	5	1.6	5.04565	*
will	initiate	6900	8	6	2.9	4.98522	*
will	take	6900	312	132	113.1	4.95154	*
could	use	215	592	13	6.7	4.92638	*
shall	execute	1134	5	2	0.3	4.9242	*
shall	strike	1134	5	2	0.3	4.9242	*
must	misuse	1684	8	3	0.7	4.89667	*
may	transfer	5113	132	47	35.4	4.88902	*
must	attack	1684	1	1	0.1	4.85128	*
must	bid	1684	1	1	0.1	4.85128	*
must	compile	1684	1	1	0.1	4.85128	*
must	conform	1684	1	1	0.1	4.85128	*
must	equip	1684	1	1	0.1	4.85128	*
must	focus	1684	1	1	0.1	4.85128	*
must	possess	1684	1	1	0.1	4.85128	*
must	pre-approve	1684	1	1	0.1	4.85128	*
must	recruit	1684	1	1	0.1	4.85128	*
must	reside	1684	1	1	0.1	4.85128	*
must	reverse-engineer	1684	1	1	0.1	4.85128	*
must	stamp	1684	1	1	0.1	4.85128	*
must	study	1684	1	1	0.1	4.85128	*
must	supervise	1684	1	1	0.1	4.85128	*
must	top	1684	1	1	0.1	4.85128	*
could	block	215	23	2	0.3	4.8357	*

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
will	provide	6900	369	154	133.7	4.8251	*
may	engage	5113	10	6	2.7	4.82148	*
would	rely	366	14	2	0.3	4.7976	*
can	open	2902	20	7	3	4.74398	*
might	put	199	26	2	0.3	4.66168	*
will	earn	6900	20	12	7.2	4.64575	*
must	conflict	1684	4	2	0.4	4.52848	*
might	mean	199	27	2	0.3	4.52554	*
should	answer	521	2	1	0.1	4.48176	*
should	detect	521	2	1	0.1	4.48176	*
should	evaluate	521	2	1	0.1	4.48176	*
should	feature	521	2	1	0.1	4.48176	*
should	lower	521	2	1	0.1	4.48176	*
will	affect	6900	106	49	38.4	4.45892	*
may	exchange	5113	8	5	2.1	4.441	*
may	introduce	5113	8	5	2.1	4.441	*
may	want	5113	8	5	2.1	4.441	*
would	result	366	60	4	1.2	4.41832	*
must	publish	1684	27	6	2.4	4.39818	*
could	mean	215	27	2	0.3	4.25721	*
can	exclude	2902	21	7	3.2	4.2385	*
may	impact	5113	13	7	3.5	4.2176	*
may	replace	5113	13	7	3.5	4.2176	*
may	entitle	5113	71	27	19.1	4.21575	*
can	grant	2902	17	6	2.6	4.14203	*
shall	debit	1134	6	2	0.4	4.13877	*
shall	indicate	1134	6	2	0.4	4.13877	*
may	commit	5113	6	4	1.6	4.13225	*
may	link	5113	6	4	1.6	4.13225	*
may	reproduce	5113	6	4	1.6	4.13225	*
will	disassociate	6900	2	2	0.7	4.06045	*
will	eliminate	6900	2	2	0.7	4.06045	*
will	lift	6900	2	2	0.7	4.06045	*
will	misappropriate	6900	2	2	0.7	4.06045	*
will	oversee	6900	2	2	0.7	4.06045	*
will	spread	6900	2	2	0.7	4.06045	*
will	strive	6900	2	2	0.7	4.06045	*
will	uphold	6900	2	2	0.7	4.06045	*
could	construe	215	5	1	0.1	4.05834	*
could	contribute	215	5	1	0.1	4.05834	*
can	demonstrate	2902	3	2	0.5	4.03718	*
can	enjoy	2902	3	2	0.5	4.03718	*
can	influence	2902	3	2	0.5	4.03718	*
can	spend	2902	3	2	0.5	4.03718	*

SLOT1	SLOT2	FS1	FS2	OBS	EXP	COLL.STR. (LOGL)	SIGNIF
will	remit	6900	5	4	1.8	4.0173	*
may	hand	5113	4	3	1.1	4.01627	*
may	launch	5113	4	3	1.1	4.01627	*
may	regard	5113	4	3	1.1	4.01627	*
can	join	2902	6	3	0.9	3.96305	*
can	offset	2902	6	3	0.9	3.96305	*
can	order	2902	6	3	0.9	3.96305	*
may	close	5113	75	28	20.1	3.932	*
will	pay	6900	292	122	105.8	3.86619	*
should	settle	521	13	2	0.4	3.84752	*