Citizen Science for Citizen Access to Law

Michael Curtotti*, Wayne Weibel+, Eric McCreath*, Nicolas Ceynowa+, Sara Frug+,Tom Bruce+
*Research School of Computer Science, Australian National University
+Legal Information Institute, Cornell University Law School

Abstract.

Over 2014, the Cornell University Legal Information Institute and the Australian National University worked with users of the Cornell LII site in a citizen science project to collect over 43,000 crowdsourced assessments of the readability of legal and other sentences. Readers ("citizen scientists") on legislative pages of the LII site were asked to rate passages from the United States Code and the Code of Federal Regulations and other texts for readability and other characteristics. They were also asked to provide information about themselves as part of the audience that uses legislation online. The overall aim of the project was to develop empirical insights into characteristics of law that may make it easy or hard to read for the audience that use it. Also, the project aimed to assess machine learning for automatically predicting readability of legal sentences at sentence level.

A major focus of this paper is to report results and insights from demographic data collected during the study. Understanding the audience which reads the law is directly relevant to readability - as the relevant question is readable by whom? Who are the citizens for whom "citizen access" might be enhanced? The paper also describes methods used to rank sentences by readability, using the data provided by citizen scientists. Finally, the paper reports initial tests on the viability of machine learning as a means of predicting readability in advance. The exploratory machine learning results reported here will be extended in further work reported in a future paper.

The research provides insight into who uses legal rules and how they do so. We draw conclusions as to the current readability of law, as well as the spread of readability among legal rules. The research creates a dataset of legal rules labelled for readability by human judges. As far as we are aware, this research project is the largest ever study of readability of regulatory language and the first research which has applied crowdsourcing to such an investigation.

Keywords: readability, legislation, legal informatics, corpus linguistics, machine learning, natural language processing, readability metrics, cloze testing, crowdsourcing, citizen science

Table of Contents

1. Introduction

> *Citizens should be able to know and understand the law that affects them. It is unfair to require them to obey it otherwise.New Zealand Law Reform Commission & Office of Parliamentary Council*(NZ, 2007)

The readability and usability of law has long attracted critical attention from users, providers, researchers and others. This paper reports research which seeks to strengthen the empirical foundations for assessing the reading difficulty of legal rules with the ultimate aim of enhancing "citizen access" to law.

In 2013 the UK Parliamentary Counsel observed:

> *Legislation affects us all. And increasingly, legislation is being searched for, read and used by a broad range of people. It is no longer confined to professional libraries; websites like legislation.gov.uk have made it accessible to everyone. So the digital age has made it easier for people to find the law of the land; but once they have found it, they may be baffled. The law is regarded by its users as intricate and intimidating.*(OPC-UK, 2013)

In 1992 it could be said that only 'a lunatic fringe' in the public would read legislation.(Krongold, 1992) Whether or not true then, by 2013, the UK Parliamentary Counsel could confidently state that it was no longer necessarily the case that readers of legislation were legally qualified. They report an audience of two million unique visitors per month for the legislation.gov.uk site.(OPC-UK, 2013)

Most of this paper discusses a project which applies "citizen science" to the problem of making law more readable. Two sub-problems in particular are addressed, building on the crowdsourced data collected for this research project. What are the characteristics of the audience which reads the law? Which parts of legal language are difficult for its readers? Both these sub-problems are empirical in nature. Much work - including empirical work - has been done in the past with legal language (for example in the plain language movement). The use of crowdsourced techniques in a citizen science project has not been applied to this task, as far as we are aware.

While amateur science has a long and respectable history (for example in the field of astronomy), the recency of the phrase "citizen science" is underlined by its addition to the Oxford English Dictionary only in June 2014. The Dictionary defines it as *"scientific work undertaken by members of the general public, often in collaboration with or under the*

direction of professional scientists and scientific institutions".[1] Other definitions have been proposed, and one that approximates our own project in part is the following *"the participation of nonscientists in the process of gathering data according to specific scientific protocols and in the process of using and interpreting that data."*(Lewenstein, 2004; Wiggins and Crowston, 2011)

Our citizen science project uses "crowdsourcing", another recently invented term (Jeff Howe in 1996). The term (although definitionally contested) expresses the idea of engaging a large number of people outside an organisation to undertake a task or solve a problem, typically online (i.e. using web technologies). Like citizen science, precursors to crowdsourcing can be found well before the 21st century. The arrival of the internet has greatly amplified the opportunity for individuals and organisations to work together towards a shared goal and many crowdsourced projects are well-known. Crowdsourcing via the web has been applied in many fields, including in citizen science projects: for example classifying galaxies, folding proteins and identifying cometary dust collected in outer space. (Howe, 2006; Brabham, 2008; Doan et al., 2011; Hand, 2010; Asmolov, 2014; Poblet et al., 2014)

In the case of our study, citizen science has been used, not only to study the language of the law, but also to learn more about people who use that language, as well as their experience of that language. The research thus engages citizen scientists in research which involves learning more about themselves as well as objective characteristics the 'data out there'. This is necessary in the context of the goals involved, as any exercise in enhancing readability is only meaningful if it addresses readability in the context of the experience and needs of the audience for given written materials.

To undertake our study we prepared a corpus of around 1250 randomly selected sentences from four different collections of English language:

(a) 139 sentences drawn from graded reading materials;

(b) 112 sentences drawn from the Brown corpus of English;

(c) 500 sentences from the United States Code; and

(d) 500 sentences from the US Code of Federal Regulations.

The Brown corpus is a balanced collection of written American English and is used as a reference point for 'normal American En-

[1] Oxford English Dictionary http://public.oed.com/the-oed-today/recent-updates-to-the-oed/june-2014-update/new-words-notes-june-2014/ and http://www.oed.com/view/Entry/33513

glish'.(Francis and Kucera, 1964) The graded reading material is drawn from 'readers' for language learners.[2] This corpus represents a modified written English simplified to be accessible to readers with different levels of reading skill. Both the Brown and Graded corpora provide reference points for calibrating and validating assessments of the legislative corpora. The US Code and the Code of Federal Regulations constitute the primary subjects of study. It may be noted that legislative rules (such as those drawn from the US Code and Code of Federal Regulations) have something in common with the graded corpus. They are also a form of modified English. Although simplicity is not the primary goal of legislative drafting - clarity, simplicity and readability are subsidiary goals that the creators of legislative texts pursue and regard as important.(Bowers, 1980; of Victoria, 1990; Melham, 1993; Tanner, 2002; OPC-Australia, 2003)

To obtain human judgements about the readability and other characteristics of the test sentences described above, we created an online interface which invited readers at the LII Cornell website to become research participants in a citizen science project. Participants were asked to provide objective and subjective responses to the test sentences. They were also asked to provide broad demographic information about themselves. Participants were, in particular, visitors who had browsed to a section, regulation or rule page of the US primary or secondary legislation at the LII Cornell site. The research participants are therefore the readers of legislative rules within the US context (i.e. the audience for whom readability of online legislative material is relevant).

Each participant was presented with a test sentence and they were asked to provide one of three alternative assessments of the test sentence.

(a) The participant might be asked to complete a likert question asking how strongly the participant agreed or disagreed with a statement as to how easy or hard the sentence was to read.

(b) Alternatively the participant would be presented with a cloze deletion test which asked the participant to guess up to ten missing words in the sentence.

(c) Otherwise, the participant was asked to complete a semantic differential test which asked the participant to rate the sentence on seven point scale against ten pairs of semantic opposites such as "readable-unreadable", "usable-unusable", "attractive-repulsive".

[2] Graded reader sentences extracted from graded reader passages downloaded from http://www.lextutor.ca/graded/. No longer available at time of publication. A copy of the corpus can be obtained for research purposes by contacting the authors

If they wished to do so, participants could assess multiple sentences, until opting out of the study. Also participants were provided with the option of providing demographic data. This included information about their gender, age, linguistic background, place of birth, educational attainment and professional background.

In addition to the foregoing, Google Analytics data on usage of LII legislation pages was also collected and analysed.

Each sentence was rated for its "language difficulty" by combining user ratings using principal components analysis and other methods. Principal components analysis is a mathematically robust method for combining many variables about an instance of data into a smaller number of variables.

This made it possible to order the sentences by language difficulty and assign them to "easy" or "hard" classifications for later use in machine learning. Natural language characteristics (such as sentence length, parts of speech and type to token ratios) were extracted from the test sentences themselves. These features were used in preliminary machine learning tests to examine how accurate machine learning would be in predicting the assigned classes.

Some of our key results are described below. For people who read legislation online, our results included the following.

(a) On the LII Cornell site, a very small proportion of the US Code is read very often, while the bulk of the Code is read very rarely.

(b) Among our research participants, legal professionals (including law students) were a minority.

(c) In proportional terms women, those without tertiary education and Spanish speakers are under-represented among those who participated in the study.

(d) The law was easier to read for legal professionals and law students than for other others who participated in the research.

For readability, our results include the following.

(a) The project demonstrates the feasibility of long-term collection of online assessments of the readability of legal texts.

(b) From user assessments provided, we were able to rank approximately 1000 legislative sentences by language difficulty.

(c) In initial application of machine learning algorithms overall accuracy (while not very high) exceeded accuracy of traditional readability metrics.

We draw a number of conclusions from our results. It is already known that the *direct* audience of legislative materials now extends far beyond lawyers and the legally trained. The results of our study are interesting in providing a quantitative indication of the modern online audience for legislation. That the non-legally trained were the majority of respondents in our study is significant. It provides quantitative validation that non-lawyers are a *substantial* audience for legislative materials. It suggests that they may now be a majority among readers of such materials. As this result may have other explanations, further studies will be required (including on other sites) to determine whether this is in fact the case. The under-representation of women among research participants is also interesting. Again it may have a variety of explanations and merits further study. The under-representation of those without tertiary education and spanish speakers is a result that might be expected, but in this case points to the relevance of asking questions about citizen access, as a likely reason is that the under-representation is a marker for lack of access.

The result that the law is easier for lawyers than non-lawyers is not surprising. Traditionally, the law has been written by lawyers, for lawyers. It is interesting however to be able to quantify the difference. In cloze deletion tests, the legally trained outperformed the non-legally trained on legal, but not non-legal, sentences. The difference was significant, but the effect size was small. Using traditional cloze deletion test analysis, the results suggest legal language is hard for all audiences (including the legally trained). For members of the public the difficulty level was 'frustrational'.

It is interesting to note the wide spread of readability in legal sentences. This suggests that there is no inherent reason why legislative sentences must be difficult. Many legislative sentences are not. For machine learning, our results confirm for the legislative field that readability metrics can readily be improved on. Results are nonetheless preliminary and we intend to extend analysis in a future paper. We leave further discussion of results to the conclusions.

Section 2 discusses related research and theoretical frameworks. Section 3 provides an overview of the study and how it was carried out. Section 4 discusses demographic data. Section 5 discusses the methods used to rank and classify sentences for language difficulty. Section 6 discusses results obtained from initial exploratory application of machine learning.

2. Related Work

The subsections which follow provide a background to our research. Given the multidisciplinary nature of our work, a number of fields from law, research methods, statistics and computer science are relevant. The fields we address are access to law; readability; plain language; readability applied to legislation; citizen science; crowdsourced research on readability; natural language processing; machine learning; assessing the reading difficulty of sentences; likert testing; cloze testing and semantic differentials. Necessarily the coverage of any particular area is as brief as possible. Nonetheless, the aggregate discourse is quite long and readers who are already familiar with these fields may wish to skip all or part of this discussion and go to Section 3 and following which describes our study and results.

2.1. Access to Law

Access to law has a number of possible meanings. The New Zealand Law Commission and the New Zealand Parliamentary Counsel's Office identify three.[3] Firstly, access in the sense of 'availability' to the public (such as via hard copy or electronic access). Second, 'navigability' - the ability to know of and reach the relevant legal principle. Finally, 'understandability' - that *'the law, once found, [is] understandable to the user.'* (NZ, 2008) We are primarily concerned with access to law in this third sense.

In 1983, a Parliamentary draftsman, F.A.R. Bennion, observed: *"It is strange that free societies should ... arrive at a situation where their members are governed from cradle to grave by texts they cannot comprehend."* The startling character of this observation arises from an incongruity of notions of 'freedom' and 'democracy', with the reality that most members of society are unable to access the meaning of laws which set out their rights and responsibilities as citizens. Ironically, Bennion himself believed that laws were written for lawyers and legal professionals and nothing could really be done about it.(Curtotti and McCreath, 2012)

This is not a view that is widely held and a number of sound democratic and other reasons have been advanced as to why laws should be understandable by all those to whom they are addressed.

[3] While our study was conducted on an American legal website using American legal text, the case for greater readability of legal materials is general across the english speaking world, and indeed beyond. Accordingly, our discussion draws on the most helpful materials. wherever we have found them.

Arguments from Rule of Law: One argument is based on the rule of law. If laws cannot be understood, it becomes difficult to sustain the rule of law, as the laws themselves are inaccessible. Implicit in this rationale is that the rule of law is in itself a social good: a social good which is frustrated by poor communication.

Arguments from Equity: Another argument is based on fairness: that to expect citizens to obey rules they cannot understand is unfair.

Arguments from Legislative Effectiveness: From the viewpoint of the legislator, adopting laws which cannot be understood is inefficient, at best, or futile, at worst. The legislator presumably wishes to communicate so as to optimally achieve its intent.

Arguments from Economic Efficiency: From the viewpoint of economic efficiency, the language should result in minimal regulatory burden. Efforts at tax law simplification are of this kind. Beyond preserving resources for other uses, implicit in this kind of reasoning is that freedom is a social good - limitations of which should only be imposed to the extent necessary to achieve a regulatory intent.

Arguments from Audience: As the Good Law initiative notes, the audience of legal rules has changed. Laws are available on the web and they are read by everyone. Laws should be written for the audience which reads it. Implicit in this rationale is a customer or citizen service orientation. Law is a service provided to its end 'users' and should be optimally *designed* to meet the needs of its users.[4]

Arguments from the Commons: A 'commons' argument regards the law as a form of property which in a sense 'belongs' to everyone. This principle underlies the founding documents of the Free Access to Law Movement. The Declaration on Free Access to Law states: *"Public legal information ... is part of the common heritage of humanity ... [it] is digital common property and should be accessible to all on a non-profit basis and free of charge."*[5]

Arguments from Rights: Close to the commons argument are rights arguments. Some authors argue that there is, or should be a 'right to access the law'.

[4] Note that the demographic results that we describe below provide an empirical description of the user base of the US legislative material.

[5] http://www.worldlii.org/worldlii/declaration/

Arguments from Democracy: As implicit in Bennion's observation cited above, open access to law can also be argued from a democratic viewpoint.[6]

Of course, these arguments apply to access to law in all three of its senses. For example, the Free Access to Law Movement began with a focus on access to law in the sense of universal free online availability.

Those who create the law are well aware of the need for it to be as accessible as possible. The Australian Office of Parliamentary Counsel put it this way in its plain language drafting guidance.

> *We also have a very important duty to do what we can to make laws easy to understand. If laws are hard to understand, they lead to administrative and legal costs, contempt of the law and criticism of our Office.* (OPC-Australia, 2003)

2.2. What is readability and how is it measured

DuBay reviews a number of the definitions that are offered for readability: *'readability is what makes some texts easier to understand than others'*; *'the ease of understanding or comprehension due to the style of writing'*; *'ease of reading words and sentences' as an element of clarity*; *'the degree to which a given class of people find certain reading matter compelling and comprehensible'*; and *'The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it.*(DuBay, 2004)

From the early 20th century researchers of language began to develop ways to measure the readability of language. A variety of "readability metrics" were developed. Such measures were used by educators to rank material for appropriate age levels. Writers also used the metrics to make their writings more usable for their intended audiences.(DuBay, 2004)

Reading measures such as the Flesch, Flesch-Kincaid, Gunning, Dale-Chall, Coleman-Liau and Gary-Leary are among the more than 200 formulas which have been developed to measure the readability of text. These formulas (although varying in formulation) address two underlying predictors of reading difficulty: semantic content (i.e. the vocabulary) and syntactic structure. Vocabulary frequency lists and sentence length studies both made early contributions to the developments of formulas. The Flesch formula calculates a score for reading

[6] For a more detailed description of the principles of access to law discussed above see: (Curtotti and McCreath, 2013).

difficulty using average sentence length and average number of syllables per word. Formulas of this kind are justified on the basis of their correlation with reading test results. For example, the Flesch formula correlated at levels of 0.7 and 0.64 in different studies carried out in 1925 and 1950 with standardised graded texts.(DuBay, 2004)

Most work undertaken on readability assesses passages of a given length (often 100 or more words). This arose because most of the creators readability metrics were seeking to use them to rate passages for inclusion in educational materials. The approach is ill suited to identifying specific linguistic features that contribute to difficulty of legal language. In a larger passage, the metric is spread over a broader vocabulary - and over a potentially large number of syntactic constructs. Greater resolution is required to be able to distinguish specific language elements contributing to language difficulty.

The uses and abuses of readability formulas have been widely debated. Readability metrics were not conceived as ways of improving the writing of text, rather they were designed to help teachers select appropriate existing texts for children of different ages.(Woods et al., 1998)

In 1993, a report to the Australian Parliament (having reviewed use of readability metrics) expressed a lack of confidence in using readability metrics on legislation. The report commented:

> *'Testing for the readability of legislation by using a computer program is of limited value. The most effective way of testing legislation is to ask people whether they can understand it - a comprehension test.'*(Melham, 1993, p xx)

2.3. Plain language, readability and legislation

Concerns about the readability of law are far from new. In England, against the resistance of the legal profession, legal language had to be prized from the medieval but firm grip of French, Latin and technical legalese. Again in Georgian times there was a 'clamor for legible [legal] English'. Again the profession opposed reform, in that case with success.[pp 124 et seq, pp133 et seq](Mellinkoff, 1963) In the 19th century, laws of the British Parliament still consisted of great slabs of discursive text. In the early-nineteenth century, Jeremy Bentham (credited with being one of the writers influencing later reforms) vociferously critiqued the problems of legislative drafting. His critique included the failure to use such obvious tools as division of legislative texts into digestible portions and section numbering to aid retrieval.(Bowers, 1980), [pp 250-251](Bentham, 1843) Practices such as section numbering and the breaking up of text were officially endorsed with the passage of Britain's

first Acts Interpretation Act in 1850. These reforms were bedded down after the first parliamentary drafting office was established in the late nineteenth century. Such offices reformed legislative drafting, including by structuring Acts in parts and use of sub-paragraphing. (See for example [p 250](Bentham, 1843),(Evans and Jack, 1984; Renton, 1975; Bowers, 1980))

In modern times, the United States also pursued plain english in the law, building on its own history of concern about legal english. In 1963 David Mellinkoff's book *The Language of the Law* appeared with the aim of "making an existing language better perform its function". In the 1960's and 70's, plain language began to appear in some insurance and consumer contracts. In the 1970's and 80's, state and federal laws began to mandate the use of readily understandable language in legal documents.(Friman, 1994) "In June 1998, President Clinton directed all federal agencies to issue all documents and regulations in plain language."(DuBay, 2004) The Plain Writing Act of 2010 mandates that US government agencies use language the public can understand and an executive order issued by President Obama in 2011 requires regulations to be "accessible, consistent, written in plain language, and easy to understand."[7]

Now it is possible to state that legislative drafting offices frequently commit to plain language as a goal they pursue.(Kimble, 1994; OPC-Australia, 2003)

Proponents of plain language cite extensive empirical studies validating the benefits of plain language. In the research field, extensive work has been undertaken to study the effect of improving legislative language.

An early example was a study reported in 1984 in which cloze testing was undertaken on several samples of legal text including legislative language. One hundred generally highly educated non-lawyers (28% had undertaken some postgraduate training) were tested. The group averaged 39% accuracy, a result close to 'frustrational' level for cloze testing. Ten participants, who had only high school education, experienced even greater difficulty, averaging 15% - a result consistent with total incomprehension.(Benson, 1984)

In 1999, Harrison and McLaren studied the readability of consumer legislation in New Zealand, undertaking user evaluations, including cloze tests. The study found traditional readability metrics to be unreliable. The results of cloze testing extracts from the legislation led to the conclusion that the legislation would require explanation before being comprehended at adult level. For young adults (aged 18-34),

[7] Plain Language: It's the Law. http://www.plainlanguage.gov/plLaw/

comprehension levels were even lower (within the frustrational level). Participants complained of the length of sentences and most felt there was a need for some legal knowledge to understand the text. All felt the text should be made easier.(Harrison and McLaren, 1999)

In the early 1990's Australia, New Zealand and the United Kingdom pursued tax law simplification initiatives which involved rewriting at least substantial portions of tax legislation. In Australia's case cloze testing on a subset of the work was inconclusive. Participants found both the original language and the rewritten language difficult.(James and Wallschutzky, 1997) Smith et al., reviewing the effectiveness of the same program, concluded that results fell 'far short of an acceptable bench-mark'. They used the Flesch Readability Score, finding that readability of sections of tax law replaced in the tax law improvement program, improved on average from 38.44 to 46.42 - a modest improvement. Even after improvement, the legislation remained difficult to read. Over 60% of the revised legislation remained inaccessible to Australians without a university education.(Smith and Richardson, 1999)

A 2003 review of the Capital Allowances Act in the UK, which was rewritten as part of the UK's tax law improvement program, undertook interviews with a number of professional users. These professionals in general responded that the new legislation was easier to use and more understandable.(OLR, 2003)

A similar review of the Income Tax (Earnings and Pensions) Act, also carried out in the UK, again found that the interviewed group (primarily tax professionals), were largely positive about the benefits of the simplification rewrite. They expressed the view that the revised legislation was easier to use and understand, although also noting the additional costs of re-learning the legislation.(Pettigrew et al., 2006)

A 2010 study of the effects of the tax law simplification in New Zealand used cloze testing to determine whether the simplification attained its goals. They reported that most of their respondents (mainly respondents unfamiliar with the tax system) found the cloze testing either difficult or extremely difficult. They found that the older (unamended) Act was the least difficult - a finding contrary to their expectation given earlier research in New Zealand. This they attributed to the nature of the selections from the older legislation. The overall average cloze results was 34.17, with unfamiliar respondents achieving 30.86%. They note that less than 25% of their subjects were able to exceed the instructional level of 44%.(Sawyer, 2010)

A study in Canada carried out usability testing on plain language and original versions of the Employment Insurance Act. Members of the general public and expert users were recruited to carry out testing.

All respondents, particularly those from the general public, found navigation and comprehension difficult, irrespective of version. Also, for all versions respondents faced difficulty in understanding the material. These findings indicated that while plain language reduced difficulty it did not eliminate it. Nonetheless participants preferred the plain language version and found it easier to use.(GLPi and Smolenka, 2000)

Tanner carried out empirical examination of samples of Victorian legislation, assessing them in light of plain language recommendations of the Victorian Law Reform Commission made 17 years earlier. In a study of six statutes, he found that the average sentence length was almost double that the Commission recommended (i.e. an average of 25 words). Also, over time, sentence length had increased. Although he also notes some improvements, he concludes: *"The net result is that many of the provisions are likely to be inaccessible to those who should be able to understand them. This is because the provisions 'twist on, phrase within clause within clause'."*(Tanner, 2002)

An empirical study of the usability of employment legislation in South Africa found that respondent accuracy improved considerably with a plain language version of the legislation. The respondents who were drawn from year 11 school students averaged a score of 65.6% when tested on the plain language version, whereas the control group scored an average of 37.7%.(Abrahams, 2003)

The empirical readability research suggests two conclusions. Firstly writing in plain language assists comprehension of legislation. Secondly legislation is generally incomprehensible or difficult to read to large sections of the population, even in those cases where plain language revision has been undertaken.

2.4. Citizen science and crowdsourcing for assessing language difficulty

As noted in the introduction, citizen science is not new. However, the availability of the internet and software has made engaging volunteers in scientific work far easier than it was in the past. Wiggins and Crowston undertake an extensive review of citizen science projects in a number of dimensions. They identify five mutually exclusive types of projects: action, conservation, education, virtual and investigation. Action projects are focussed on engaging volunteers to address local issues. Conservation addresses natural resource management. Investigation refers to scientific investigation in a physical setting. Virtual projects have similar goals to investigation projects, but in an online setting. Education is primarily concerned with education and outreach. They also note that citizen scientists may be engaged in data collection

and analysis, participation in project design and in drawing conclusions and disseminating results. Citizen science projects are typically organised in a top down fashion by a scientific team and volunteers are recruited to assist in the conduct of the project. This is also true virtual projects which typically have a 'top down' organisation. However, sometimes citizen science are organisationally 'bottom up', though this is largely limited to local projects. Scientific issues arise for all types of citizen science. For virtual projects the primary scientific challenge is scientific validity of results and achieving a design that maintains participant interest. Success depends on reaching a critical mass of contributors. The primary approach to ensure validity is replication of results. (Wiggins and Crowston, 2011)

Citizen science projects (particularly those carried out online) can be appropriately considered a form of crowdsourcing. Crowdsourced typologies are similar to those for citizen science projects, particularly as to how the crowd is involved. Poblet et al identify a hierarchy of crowd involvement, based on the type of data that is being crowdsourced. At the base, the crowd may merely serve as sensors (as in data automatically generated by mobile devices). The crowd may be "social computer" - i.e. generators of data later available for assessment (as an indirect rather than intended outcome). The crowd may serve as reporters (i.e. information generators). The crowd may be microtaskers (i.e. performing specific tasks over raw data).(Poblet et al., 2014) Asmolov discusses a broader typology of crowdsourcing, extending the analysis to the question of the level of crowd engagement. At one end of the spectrum is full organizational control - at the other the organization is merely incidental to the crowdsourced activity. The character of crowdsourcing is also disputed: is it the wisdom of the crowd - or the crass capitalist exploitation of unpaid workers? Is it participatory or is it exploitative? Of course, different projects may have one or other of these characteristics. Key to understanding crowdsourcing is what it does: it enables action through accessing resources of the networked crowd (e.g. intellectual, computational, physical or financial).(Asmolov, 2014)

In our own study, all the citizen science dimensions discussed above are in play. Our research is firmly within the virtual space and displays the characteristics mentioned by Wiggins and Crowston for that space. Organizationally, the project was framed in a top down fashion. Citizen scientists were primarily asked to participate in assessing data on a platform designed without their involvement. The platform was designing in a way which was hoped to maintain interest; providing a variety of tests, as well as exploring different ways of assessing readability through crowdsourced evaluations. After data collection, careful

review of data was required to remove confounding inputs (i.e. data validation). Replication of input was a primary means of controlling for 'bad' data. The size of the participant base became an issue, limiting how far the project could go (i.e. how quickly data could be collected). It would be interesting to attempt to expand the scope of citizen scientist participation in future projects, though this may skew participation away from a balanced reflection of the audience for online legislative materials.

The term citizen science has not been used in the readability sphere, nonetheless there are a very small number of projects (under the rubric of crowdsourcing), which also amount to citizen science projects. We have only been able to identify two which were focussed on validating crowdsourcing as a method for readability studies.

De Clercq et al. evaluate the effectiveness of crowdsourcing as a method of assessing readability. They compared the accuracy of crowdsourced human judgements of the readability to those of expert judges, finding a high level of agreement in readability ranking between the experts and crowdsourced users. Crowdsourced users were presented with two randomly selected texts of one to two hundred words and invited to rank them by readability. Expert teachers, writers and linguists were given a more complex task of assigning a readability score to each presented text. The researchers concluded that crowd sourced user judgements and expert judgements were highly correlated as to readability ranking. They found also that readability metrics had a lower correlation with both these two judgement sets.(De Clercq et al., 2013)

A more general study by Munro et al. concluded that there was a high correlation between traditional laboratory experiments and crowdsourced based studies of the same linguistic phenomena. Among their conclusions was that crowdsourced judgements closely correlated with cloze testing results. (Munro et al., 2010)

We are unaware of any previous studies which have used crowdsourcing to assess the readability of legislative texts.

2.5. Natural language processing and machine learning

Recent years have seen a growing body of research applying natural language processing and machine learning to assessing the readability of text. The term 'natural language processing' represents the capacity of computers to hold and analyse potentially vast bodies of text. Natural language processing typically transforms natural language into collections of variables representative of the characteristics of the natural language. Such characteristics range from the raw text itself, to

representations of the syntactic and vocabulary characteristics of a text. Such characteristics are then available of further processing or analysis.

Machine learning is a well elaborated process. In summary, it seeks to make predictions based on a body of data. Characteristics from that data is extracted as 'input features' and provided to one or more of a variety of machine learning algorithms. The most common goal is for the algorithm to be able, based on patterns in the data, to return a model which predicts the class of a previously unseen item of data. Machine learning includes both 'supervised' and 'unsupervised' learning. In supervised learning, training data (already labelled with the appropriate classifications) is provided to 'train' the learning algorithm. In the unsupervised case, the machine learning algorithm tries to separate the data into natural groupings based on clusterings of features.[8]

Both natural language processing and machine learning have been applied to automatically predict readability. An exhaustive review is not carried out here but a number of aspects of particular interest are highlighted. A key question is what features might assist us in assessing readability? Studies have systematically examined sets of features for their utility in assessing readability. The most easily extracted features are readability metrics and 'surface' features such as average sentence length, average word length, average syllable length, capitalisation and punctuation. Other features studied include lexical features such as vocabulary and type/token ratio,[9] parts of speech frequencies, ratio of content words to function words, distribution of verbs according to mood, syntactic features such as parse tree depths, frequency of subordinate clauses, ngram language models, discourse features, named entity occurrences, semantic relationships between entities and anaphora occurrences.(Si and Callan, 2001; Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Heilman et al., 2008; Pitler and Nenkova, 2008; Kate et al., 2010; Aluisio et al., 2010; Feng et al., 2010; Dell'Orletta et al., 2011; Kauchak et al., 2014) A good overview of the state of the art is provided by Collins-Thompson's survey article on readability research using machine learning.(Collins-Thompson, 2014)

Applying natural language processing and machine learning to predict readability has made considerable progress over the last decade or so. Studies such as those referenced above demonstrate that prediction of readability of text can be improved by incorporating higher level

[8] See Bird et al. for an accessible and practical introduction to natural language processing. Chapter six introduces machine learning for classifying text.(Bird et al., 2009)

[9] A 'type' is, say, the word 'red' and a token is any word. So in the phrase "the cat sat on the mat" the type to token ratio is 5/6, as the word 'the' occurs twice.

linguistic features into predictive models. It is also notable that only initial steps have been taken to apply findings in this field to identifying reliable methods to *improve* readability of text.

A limitation of such methods is that without a considerable body of labelled data, it is difficult to attain high levels of accuracy with machine learning. The use of crowdsourced methods enables this problem to be addressed.

2.6. Assessing the readability of sentences

Historically, as we have seen above in section 2.2, readability has been addressed at the level of at least a passage of text. Klare notes that readability metrics are designed for larger blocks of text providing a connected discourse. They won't work well on disconnected fragments or single sentences.(Klare, 2000) Fry is one of the few who as early as 1990 sought to create a metric better suited to short passages. At the time, most metrics were designed for use with passages of 300 words or more. Fry particularly noted that such 'short passages' were important in materials such as 'science textbooks, math textbooks, passages used in tests, manufacturers' warranties, and rules and procedures in driver's training booklets.' Fry's new metric could be applied to passages with 40 words or more. However Fry stated that the new metric was only appropriate for passages with at least 3 sentences - making it inapplicable to detecting readability of single sentences.(Fry, 1990)

For this study, sentences were chosen as the unit of study. A rationale for this choice is that the sentence is the basic unit of content for legal rules. That is, in many jurisdictions, each rule is contained in a separate sentence. Another reason for this choice is that longer passages are not sufficiently granular to automatically identify features which contribute to reading difficulty. Without this level of granularity, it is difficult to automate recommendations for improving how materials are written.[10]

Studies exploring language difficulty at sentence (or smaller) level have only emerged recently; with the availability of computational tools which make it more practical.

A number of studies exist which seek to explore smaller units of text. Kanungo and Orr carry out a study of snippets of text returned as web search results which are either sentences or sub-sentences. They present a study involving 5000 human judgements of the readability of such short text fragments. They apply machine learning using a gradient boosted decision tree as the learning model. Their study assesses a number of features (e.g. fraction of capitalisation and fraction

[10] It should be noted however that semantic meaning is often connected across sentences. Analysis of this broader level of meaning is lost at sentence level.

of search terms) as predictors for the reading difficulty assigned by human judges. They also assess traditional reading metrics such as the Fog, SMOG, Flesch Kincaid metrics. They find that metrics had virtually no correlation with human judgements of the readability of search results . On the other hand, the Pearson correlation R of their boosted decision tree model correlated at around 63%. This study illustrates the inapplicability of traditional metrics to short language segments and to specialised language (i.e. search results in this case).(Kanungo and Orr, 2009)

Dell'Orletta et al note that much work on readability in the natural language processing field is focussed at document level but that such methods are unreliable at sentence level. They study readability at sentence level on the rationale that this would be useful for text simplification. (A rationale that applies in the context of enhancing access to law). They develop a model capable of accurately labelling sentences for reading difficulty with 78% accuracy. Their model includes a range of linguistic features beyond those traditionally used in readability formulas.(Dell'Orletta et al., 2011)

Sjoholm is another researcher who assesses the readability of sentences. He notes the absence of existing metrics for predicting readability at sentence level. He builds on previous studies by developing a probabilistic soft classification approach that rather than classifying a sentence as 'hard' or 'easy' gives a probability measure of membership of either class.(Sjöholm, 2012)

2.7. Likert testing

Likert testing is widely used by researchers. It is a test of a person's subjective response to a statement. Most often the test asks how strongly a person agrees or disagrees with a particular statement put to the person. A common form allows participants to select between five possible responses: 'strongly agree', 'agree', 'neither', 'disagree' and 'strongly disagree'.(Heiberger and Robbins, 2013) Figure 2 is an example of a likert test presented to participants during our study.

Likert testing has been applied to readability studies in previous research. Heydari employs likert testing to evaluate the readability of ten passages.(Heydari and Riazi, 2012) Hall and Hanna use likert testing to assess the effect of colour on readability of web pages.(Hall and Hanna, 2004) Ferrari and Short apply likert testing to evaluate the effect of size and font type on readability.(Ferrari, 2002)

Kandula et al use seven point scale likert questions with a cohort of experts and patients to rate the readability of health literature. They are concerned with the difficulty of health literature which they note

the Institute of Medicine assessed was difficult to read or act on by more than half of the US adult population. They found a high level of correlation (.81) between expert and patient ratings of language difficulty.(Kandula and Zeng-Treitler, 2008)

The appropriate analysis of likert scale items is a matter of controversy among researchers and the question is relevant to analysis of our results. Different camps argue for different analysis methods. Essentially, the controversy concerns whether parametric as opposed to non-parametric tests can be used to analyse likert data.[11] Clason et al argue that likert items must always be treated as ordinal, even when combined in a scale, and therefore argue for non-parametric testing. (Clason and Dormody, 1994) Norman critiques such arguments, arguing that parametric tests are often robust even when assumptions (such as normality) are violated. *"[B]oth theory and data converge on the conclusion that parametric methods examining differences between means, for sample sizes greater than 5, do not require the assumption of normality, and will yield nearly correct answers even for manifestly non-normal and asymmetric distributions like exponentials."* Norman concludes that parametric tests are appropriate for analysis of likert data both for differences of means and correlation of data.(Norman, 2010) Similarly de Winters et al, who undertook a systematic comparison of t-tests (a parametric test) and the Mann Witney Wilcoxon test (a non-parametric test) on a diverse range of distributions of data concluded that the differences between the tests was minor and exceeded 10% only for a few of the 98 distributions they studied.(de Winter and Dodou, 2010) To the extent that parametric analysis is used on likert tests in this paper, there is sufficient support for it in the research literature.

2.8. Cloze testing

The cloze procedure involves testing the ability of readers to correctly re-insert words that have been deleted from a given text. Typically the test is administered by deleting every nth word in a text. When used to assess the readability of a text, the cloze procedure is administered by deleting every fifth word (including sometimes five different versions of the text staggering the deletion), and replacing it with a blank space. The reader must fill in the missing terms.(Bormuth, 1967) Figure 3 is an example of a cloze test used in our research.

Although initially conceived as a remedy for the shortcomings of readability formulas, the cloze procedure came to complement conven-

[11] Parametric tests (such as the students t-test and ANOVA testing) make more assumptions about the test data than do non-parametric tests.

tional reading tests.(DuBay, 2004) Cloze procedure was also developed to provide a more valid measure of comprehension than traditional multiple choice comprehension tests.(Wagner, 1986) Of greatest interest in this context is use of cloze tests as a measure of the readability of a text. Bormuth notes that there is a high correlation between cloze readability testing and comprehension testing on human subjects:

> *A reasonably substantial amount of research has accumulated showing that cloze readability test difficulties correspond closely to the difficulties of passages measured by other methods.* (Bormuth, 1967)

Bormuth cites studies, including his own, which show correlations ranged from .91 to .96 with the difficulty of texts assessed with traditional comprehension tests.(Bormuth, 1967) When properly applied, the cloze test provides an indicator of how difficult a text was for given readers. A cloze score (i.e. proportion of correct responses) below 35% indicates reader frustration, between 35% and 49% is 'instructional' (the reader requires assistance to comprehend the material) and 50% or above indicates independent reader comprehension.(Wagner, 1986)

2.9. Semantic differentials

Semantic differentials were originally developed by Osgood in the 1950's. A semantic differential is comprised of two bipolar adjectives ('readable-unreadable' for example) with a scale in between. The research participant is asked to select a point on the scale which they consider best corresponds to the test stimulus. Typically, the user is presented with multiple semantic pairs and asked to assess a test item for each pair. Figure 4 provides an example of a semantic differential test used in our study.

Semantic differentials may vary by number of points on the scale or presence or form of labelling of scale point. A scale varies from a positive to negative end and thus has both direction and magnitude. (Garland, 1990; Johnson, 2012) Semantic differentials have been widely applied and are seen as an accurate measure of individuals 'affective' responses to a stimulus. Osgood found that users ratings of semantic differentials could be reliably grouped into three major dimensions which he labelled evaluation, potency and action. The method has been used to test individuals responses to words, pictures, facial expressions and a wide variety of concepts.(Johnson, 2012)

Garland compares three different forms of the semantic differential test to test whether the form of the test affects user responses. The three forms were: semantic differentials without labels, semantic differentials with numeric labels and semantic differentials with text labels (such as 'very', 'quite', 'neither' etc). Garland asked users to rate the test for

preference, ease of expressing opinion and ease of completion, finding that users preferred semantic differentials with text labelling. Garland also found that there was no difference in the distribution of responses of the administered semantic differential tests and concludes that the form of test used is unlikely to influence users responses. Garland does however note that numerical scales may be favoured by users who are used to working with numbers. (Garland, 1990) For these reasons, and given that 'used to using numbers' is not a characteristic applying particularly to the users of legislation we use a labelled semantic differential test in our study.

Semantic differentials have also been used to measure "user experience". User experience has been defined as "a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service". The concept is broader than concepts of usability which are more specifically concerned with functional characteristics of the artefact being tested. (Vermeeren et al., 2010) The use of semantic differentials in our study enabled a broader examination of how users responded - for example including responses to concepts such as 'leniency' or 'attraction-repulsion'.

A possible alternative to a semantic differential (in our case participants were asked to select an appropriate radio button) is the 'visual analog scale' or a slider. However, Couper et al find no advantages to use of a visual analog scale. Rather they found that using a slider led to higher levels of missing data, and longer completion times.(Couper et al., 2006)

2.10. Principal components analysis & factor analysis

In our study we collect not only multiple ratings, but also multiple ratings of multiple variables for each sentence used in the study. We need to combine the data from each of cloze, likert and semantic differential test to provide a single variable which is representative of reading difficulty of a particular sentence.

Principal Components Analysis is particularly suited to this task. Its goal is to reduce the number of dimensions in a set of observations by combining variables into a reduced number of variables.(Härdle and Simar, 2003, p 234, 241) Factor analysis similarly seeks to identify underlying latent variables (factors) by grouping together variance in the most highly correlated variables into a reduced number of factors.(Floyd and Widaman, 1995) A question for both methods is how to decide how many variables to retain after principal components or factors are extracted. One widely supported method is graphical. It looks for a bend in a curve known as a scree plot. The y-axis of the

scree plot shows eigenvalues extracted for each principal component, while the x-axis shows the extracted components themselves. Figure 16 is an example of a scree plot. Principal components to the left of the 'bend' in the scree plot are often retained. This method is regarded as among the most sound.(Costello and Osborne, 2005)

3. Description of the Study and Observations

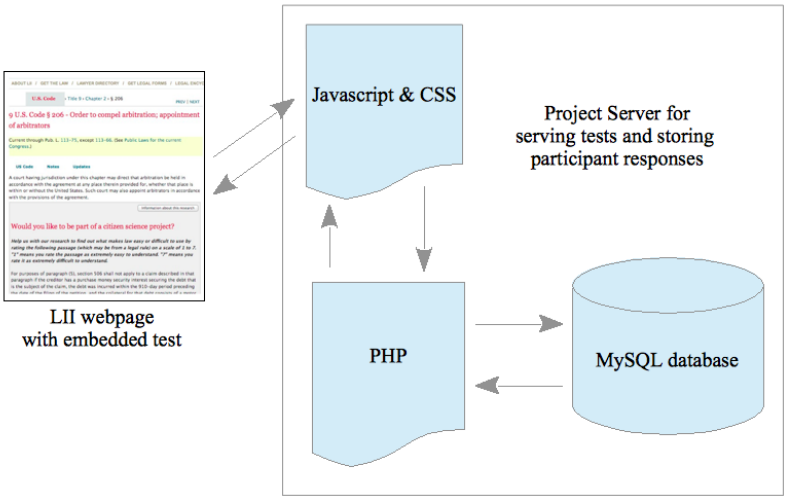

Figure 1. Project platform design

The primary data for the research was collected from 2 May 2014 until 31 July 2014 using the crowdsourcing methods described above. In total, 63,250 submissions were received from users spread across four sets of data: demographic data, likert submissions, cloze results and semantic differentials. From among these, some submissions were null results (e.g. a user pressed the submit button without providing any data). Also some data was removed as outliers for particular tests.[12] Table I below shows totals and percentages of usable data after filtering. More than 43,355 usable readability assessments were collected.

[12] For semantic differential tests, 'donkey votes' were removed - i.e. votes in which the user selected only the same value down a column. Also results where more than 30% of a semantic differential were null were removed. For cloze testing, an issue where a score of 0 was obtained, was how to distinguish genuine attempts from 'careless' input. Results were filtered if 30% or less of fields of a cloze submission had any input (i.e. an attempt at guessing the word). As noted above, the issue of data validation is a characteristic of citizen science projects.

Table I. Submissions

Database	**Total Submits**	**% of Usable Data**
Demographic Data	14912	> 93.8
Likert Data	23402	99.9
Cloze Data	12970	85.2
Semantic Differential Data	11966	> 74.6

Over the period of three months that data was collected, the rate of data collection remained essentially linear. This points to the feasibility of longer term data collection of user experience in online legal publishing environments. Also, as there was an equal chance of being asked to complete a likert, cloze or semantic differential test, the response rate for each question type is informative. Semantic differential tests were least likely to be responded to, while likert tests were responded to at almost twice the rate. As rate of data collection was an important consideration, these differences are relevant to future research design.

The tools used to undertake the research included the python programming language, used for scripts for preparing corpora and undertaking data cleaning extraction, the Weka Data Mining Software package (Hall et al., 2009) for machine learning, the R-statistical package and associated R-Cmdr graphical user interface (R-Core-Team et al., 2012; Fox, 2005) for undertaking statistical analysis, the Readability Research Platform (Curtotti and McCreath, 2013) and the Natural Language Toolkit (Bird et al., 2009) for carrying out metrics extraction and natural language processing.

The platform to enable data collection had a number of functional elements:

(a) a background server for serving test sentences and receiving and storing participant responses;

(b) php scripts which communicated asynchronously with a mysql database on the server;

(c) javascript and css files which communicated with the primary Cornell pages and with the php scripts.

Figure 1 provides an overview of the platform design. Brief code snippets in the primary LII pages linked the platform with the web pages viewed by participants.

Figure 2 illustrates likert tests used in our study. The participant was presented with a sentence selected from the four test corpora and

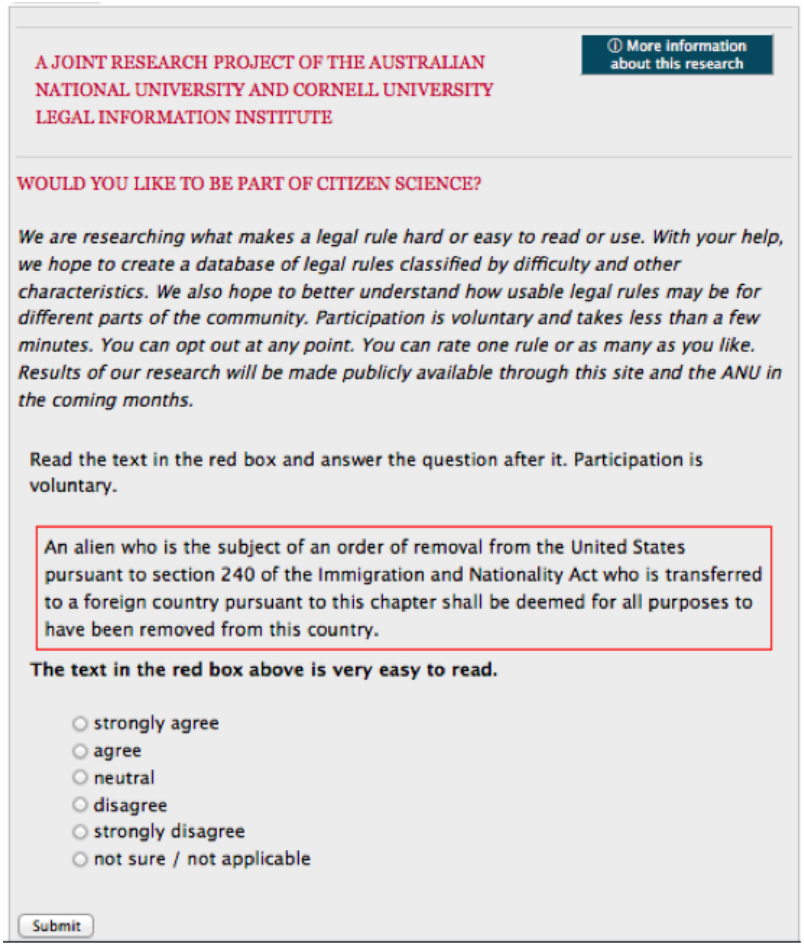

Figure 2. An example likert test presented to research participants

was asked to indicate their level of agreement from "strongly agree" to "strongly disagree" with a statement about a sentence. That statement could be that the sentence was "very easy", "easy", "hard" or "very hard" to read. For example the user might be presented with the statement that "The text is very easy to read" and asked to indicate their level of agreement with the statement.

Figure 3 illustrates a cloze test. In this case, the participant was asked to guess up to ten missing terms in a test sentence.

Figure 4 illustrates a semantic differential test. Here, the participant was asked to rate a test sentence against each of ten semantic differentials. As we wished to minimise potential disruption to the normal use of the LII website, tests were presented at the bottom of LII pages.

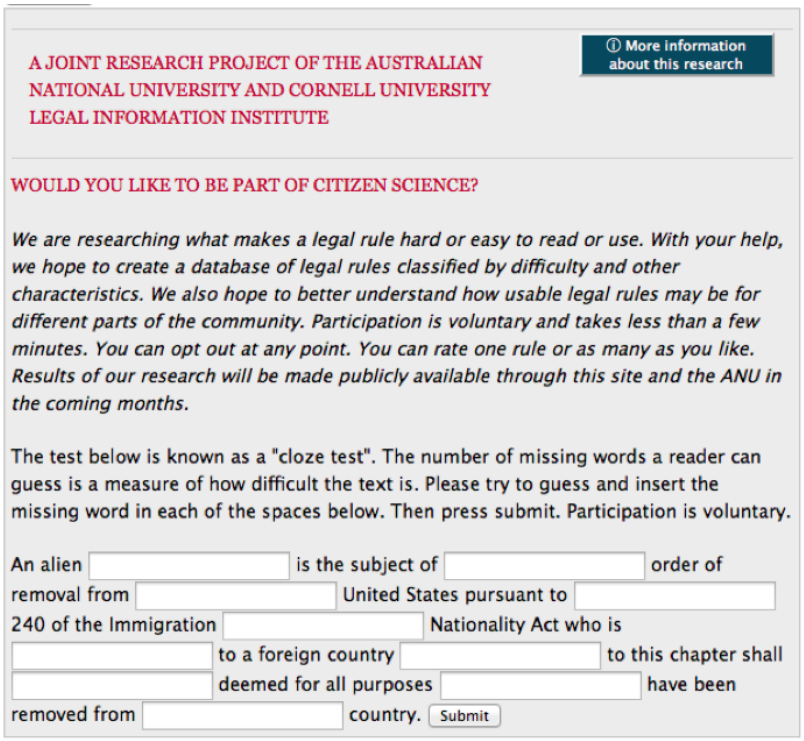

Figure 3. An example cloze test presented to research participants

4. Demographics

The demographic results from our research is a particular focus of this paper. We first present Google Analytics from the LII site which provides an independent source of data addressing user behaviour on the site. We then discuss the demographic data collected during our study.

4.1. What law do people read? Insights from google analytics

Google Analytics were studied for visits on the Cornell LII legislation pages over a period of 12 months. In the period 18 October 2012 to 17 October 2013 a total of 927.4 person years were spent reading legal rules at the LII site (this includes the US Code, CFR, UCC, constitution, rules of procedure etc). Most people found their way to legislative provisions directly by searching for the relevant legal rule (i.e. the landing page on the LII server was a specific section or regulation). This implies that often people have had some introduction to what might be relevant

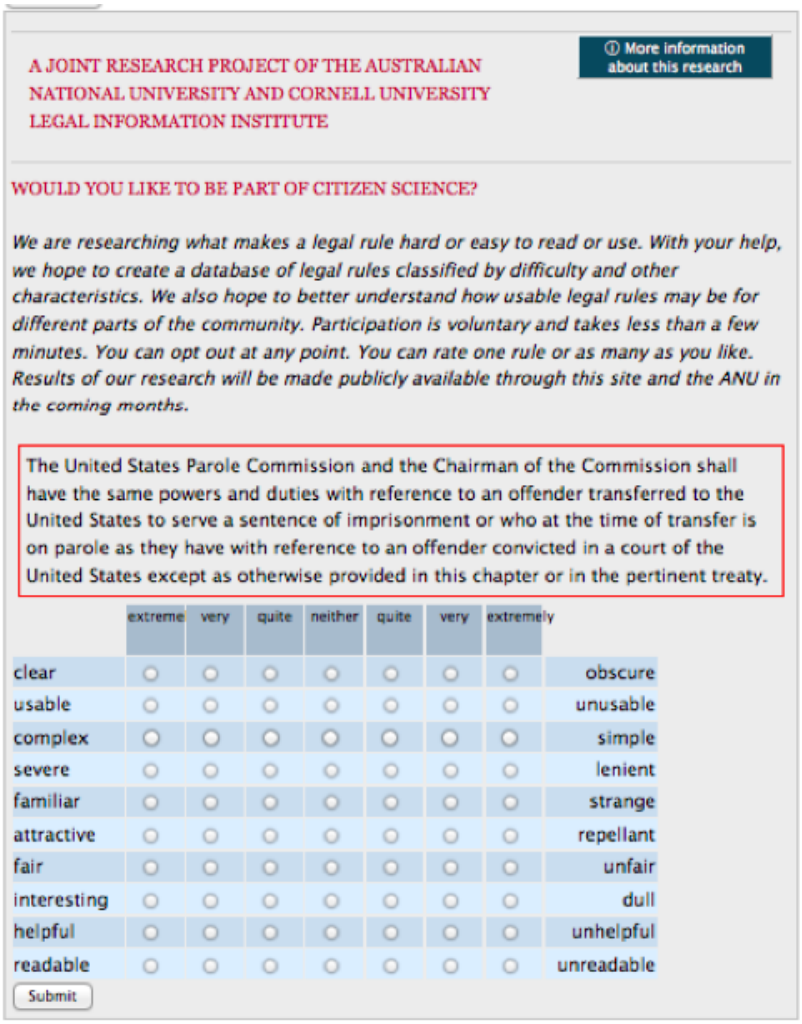

Figure 4. An example semantic differential test presented to research participants

laws for their concerns, before navigating to the LII site. The site has a large traffic, with 112 million page views during a year (of which 38.8 million page views are from US Code and 19.5 million page views are from the Code of Federal Regulations) (21 August 2013 - 20 August 2014). By comparison the official UK legislative site receives 5 million page views per week.(Tullo, 2013)

Most interesting from the readership data was its power law distribution. Far from readership of sections being equally distributed - the readers for a particular section might attract varied by many orders of magnitude. A mere 37 sections of the US Code (landing pages), for example, account for 9.97% of entire traffic to US Code sections.

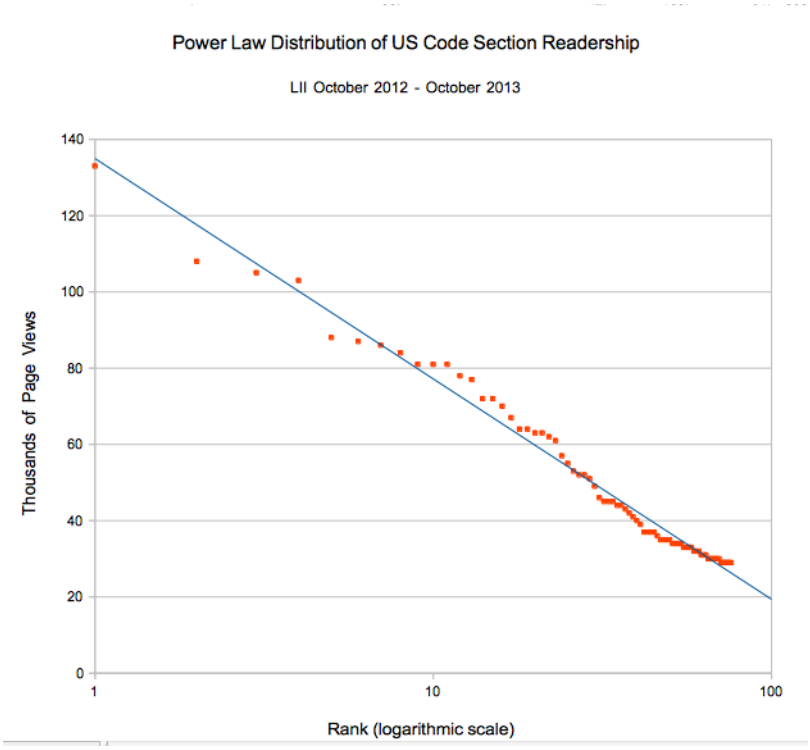

Figure 5. Power law distribution of section readership for most frequently visited sections

This was from a total 'node' count of close to 65,000 sections. Of these, 8391 sections (pages) were visited once in a 12 month period, 4267 were visited twice, and 2833 were visited 3 times. The most frequently visited section was visited 133 438 times during the twelve month period (Title 28 section 1332).

The implications are significant for the task of enhancing access to law. For practical purposes, most of the US code is of marginal relevance. It is rarely read and efforts to improve its readability may not be warranted. On the other hand, language difficulty in highly read parts of the code will impact significantly on access to law and on the regulatory burden faced by users. For a much smaller effort than full review of an area of law, a disproportionate improvement in user experience is available by addressing readability of the most read parts of the code.

4.2. Demographic data

As mentioned above, research participants were asked to provide general demographic information, but could opt out if they wished to do so. The population from which demographic data (and other data) was drawn was limited to visitors to the site who were engaging with legislative or regulatory materials (i.e. legal rules). Data was collected on age, birthplace, education, gender, language and persona. The 'persona' category refers to certain typical users of legal data: e.g. legal professionals and members of the public.

There may be systematic effects in those who voluntarily chose to participate in the study, nonetheless the results provide an indicator of the user base for online legislative information.

On questions of readability, Dubay notes the two most important questions are"the reading skills of the audience" and the "readability of the text".(DuBay, 2004) For example, if all readers of law are judges (i.e. an audience highly familiar with reading and comprehending legislative texts) readability issues will play out quite differently to a case where a substantial proportion of readers are not legally trained. In the latter case, such readers may find the language difficult and unfamiliar, and a case may be established for improving the writing of legislative texts.

4.3. Who reads the law online and why they do so

Participants were asked to nominate a broad persona that best matched the reason they used the law. The use of personas to study readability in a legislative context was described in a study of the users of UK Legislation reported by Carol Tullo of the UK National Archives Office at 2013 Law Via the Internet Conference. In the case of that research, the personas were: a compliance officer; a law librarian; a member of the public seeking to defend her rights; and a parliamentarian. It was noted that such categories do not necessarily capture the entire user base.(Tullo, 2013)

In our study five personas were used: legal professionals (including law students); non-lawyers engaged in compliance; members of the public seeking information on their rights; individuals engaged in law reform or law making; and "others". As would be anticipated legal professionals and legal students (i.e. the legally trained) were the largest single group of respondents (41.7%). However, surprisingly, they were the minority of respondents.

Members of the public seeking information on their rights (23%) and non-lawyers engaged in compliance management (13.4%) also represented substantial categories. A large "other" category represented

(18.59%) responses. This category was almost one fifth of respondents, although it is not immediately obvious what this 'other' category may represent. Those engaged in reforming the law (participants relevant to the democratic process) represent 3.5%. (See Figure 6) Meeting the needs of users drawn from the public is most directly related to "access" for reasons of equity. The compliance category represents considerations of economic efficiency. The reform category is related to rule of law and the democratic process.

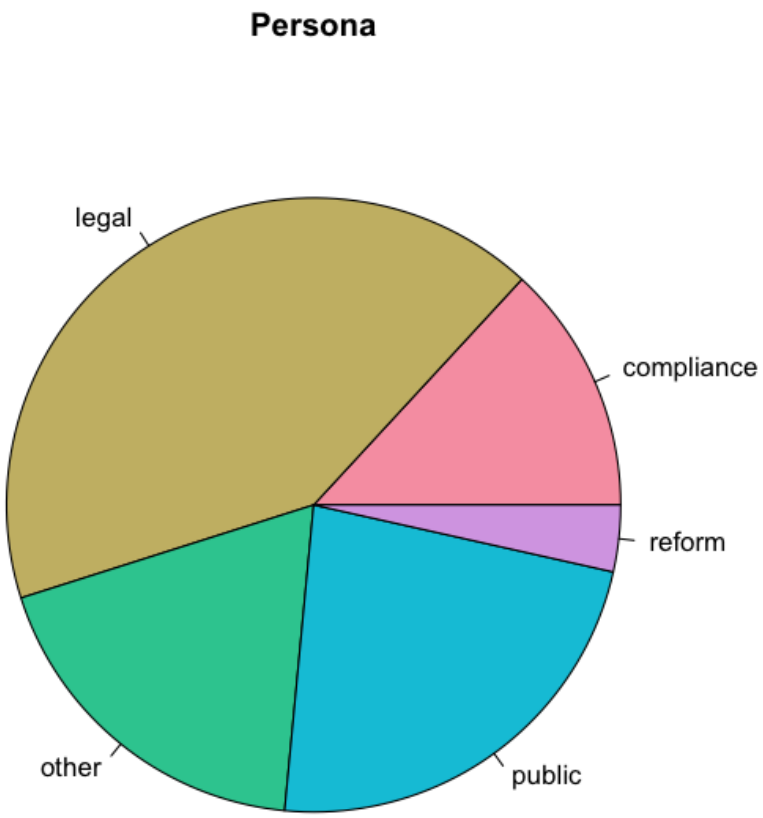

Figure 6. Who reads legislation

We also explored differences in *which parts* of the law different user categories were most likely to be reading. This data was derived from the landing or source page from which a reader participated in the research.

Lawyers were far more strongly represented in the audience for Federal Rules as opposed to the other two bodies of legislation (66% versus around 40% for the US Code or Code of Federal Regulations).

Those concerned with compliance represented 21% of the audience for the Code of Federal Regulations, dropping to 9% of the audience for the US Code. By contrast the public audience for the US Code was 27% with the 'other' group representing another 20%. Table II sets out additional results.

Table II. Percentage of Audience By Legal Code

code	compliance	legal	other	public	reform	count
Code of Federal Regulations	21.3	40.4	17.7	17.5	3.2	5119
Federal Rules	3.3	66.7	8.2	18.0	3.9	672
US Code	8.9	40.6	20.0	26.8	3.7	8188

The top eleven most frequent titles of the US Code were also examined for distribution of audience. Figure 7 illustrates the distribution of audiences. Legal professionals constituted a majority of the audience only in the cases of Title 28 (the judiciary), Title 11 (bankruptcy) and Title 36 (patriotic observances). The public were the highest users of Title 18 (crime) and Title 10 (armed forces). Individuals interested in legal reform were most highly represented in Title 17 (copyright), although in no case representing more than a few percent of total audience. Compliance officers were most highly represented in Title 26 (internal revenue). In terms of access again implications can be drawn. Criminal law is an area where the public needs the law to be readable. They are a substantial audience for the criminal law. It will be noted from Figure 7 that it is a heavily read title. The internal revenue code (unsurprisingly) represents an area where concerns relating to regulatory burden are more pertinent.

The results discussed above are consistent with the observations of the UK Parliamentary Counsel's office that *"increasingly, legislation is being searched for, read and used by a broad range of people...; websites like legislation.gov.uk have made it accessible to everyone."*(OPC-UK, 2013) The consistency of our results with the description of the user base for a major national legislative site support a conclusion that the patterns observed on the LII site are not an artefact of either the LII site or the study design. The results support a conclusion that whatever may have been the situation in the past, legal professionals are far from the primary readers of legislation in the online environment. Indeed substantial non-lawyers are a substantial audience for the law online. Some caution is required in interpreting this result, as it is possible that other reasons explain it (e.g. lawyers might have responded at a lower rate than non-lawyers). Studies on other online sites would be required

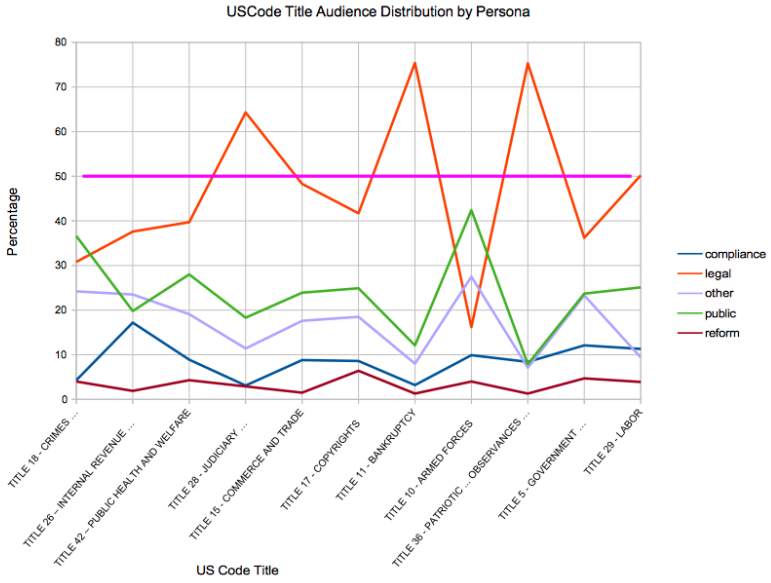

Figure 7. Persona distribution by title

to clarify whether this result generalizes to the underlying audience for law online.

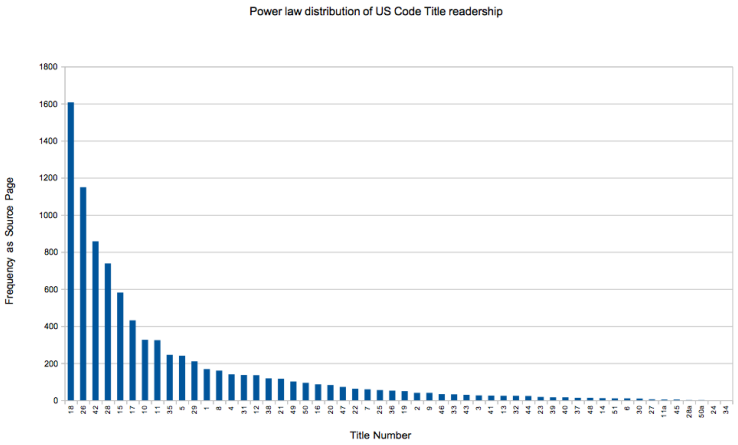

Figure 8. Relative readership of US Code titles

In light of such findings as to the user base for legal rules, "who" legal rules are being written for is practically as well as theoretically important. To write only, or primarily, for judges and lawyers fails to address the needs of a substantial proportion of users of legal rules. Also it is possible to differentiate between different parts of legislation

by level of audience interest, again carrying implications for how law might be written.

4.4. Gender results

There was a sharp disparity in the number of responses received from males and females. Women represented only 35.4% of the responses by gender. There are a number of possible explanations for this result. One possibility is that there is a gender disparity in access to law reflecting societal conditions. There may be factors in the way that law is provided online that affects its accessibility to women. Alternatively, the result may be wholly or partially an artefact of the study design. To promote participation, the study was described as an invitation to participate in 'citizen science'. If the 'scientific' description is a cause of the lower participation by women, it is a marker of gender exclusion in another social dimension. A further possibility is that the results at this site are not representative of broader usage patterns.

The legal profession, like many professions has only partially achieved gender equality. It would be expected therefore that there would be a lower representation of women in the legal profession persona and this is reflected in our results (60% to 40%). However, the gender disparity in participation is even more marked for the non-professional personas (i.e. public and other), where women represented only 30.3% and 33.5%, respectively. The differences between these different personas tend to support a conclusion that the difference in participation is 'real', rather than related to the study design. Such gender disparities merit further investigation.

4.5. Age

Users were asked to nominate an age category (grouped into 15 year age bands). The responses show a broad distribution across age groups. Figure 9 illustrates these results. However age is not evenly distributed across user groups. Legal professionals are dominated (as would be expected) by working life adults. They are also generally younger as compared with members of the public accessing law online. This may reflect the inclusion of law students in this group.

4.6. Birthplace

Users were asked to nominate a broad region of the world in which they were born. Over 85% were born in the United States or Canada. This is to be expected for a US legislative site. The overseas born population reported in 2010 for US population was 12.9%.(Grieco et al.,

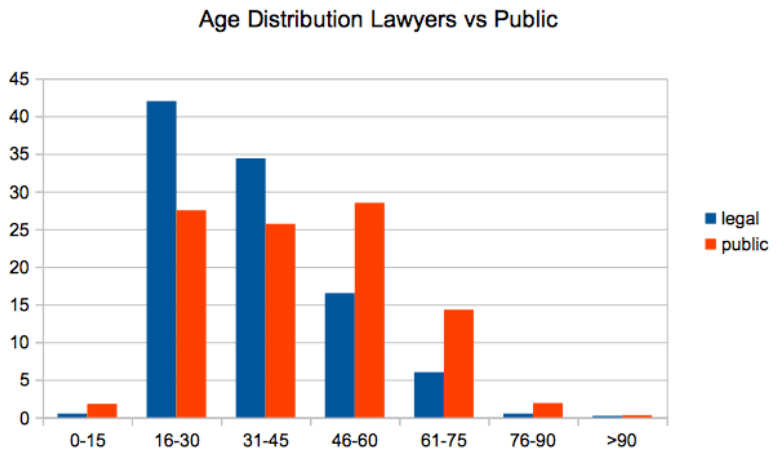

Figure 9. Legal vs public age distribution

2012) This is a similar proportion to the proportion accessing the LII legislative pages, although direct parallels are invalid, because LII users also include an unknown number of users from overseas.

4.7. Education

Users were asked to indicate the highest level of education they had completed: primary, secondary, vocational or tertiary. Overall tertiary respondents represented 78.24% of respondents. Primary were 3.74%, secondary were 9.31% and the vocationally educated were 8.72%.

As with age, educational completion varied significantly between different personas. The tertiary educated strongly dominated the legal profession (94.3%), as would be expected. For the public, the proportion of tertiary educated was 56.9%. This figure is considerably higher than the completion of tertiary education in the US population as a whole. Again this suggests an access issue. Those with primary, secondary and vocational education are under-represented among readers of law. In 2009, between 20% and 30% of the population over age 25 had completed tertiary education.(Ryan and Siebens, 2012)

Again this carries implication for access to law. Many without tertiary education may not even be attempting to read the law. In addition, 43.1% of those who were among participants did not have tertiary education. To address the needs of the public, the law needs to be written to take account of the fact that a substantial proportion of its readership does not have tertiary education. In terms of the regulatory burden, of those concerned with compliance, 16.6% do not have tertiary

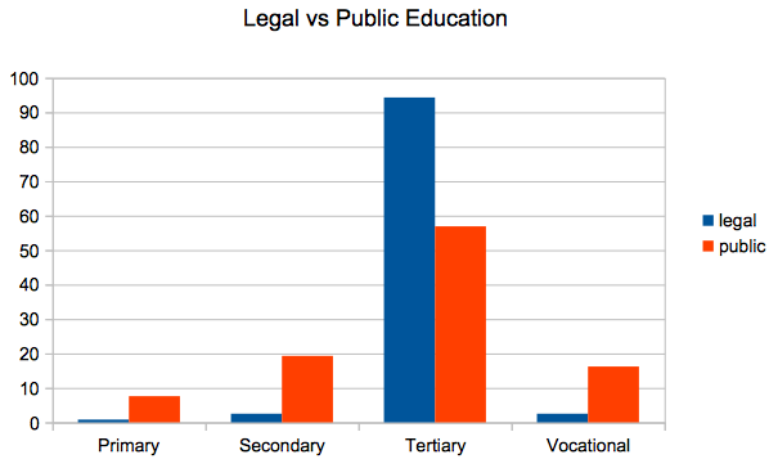

Figure 10. Legal vs public educational attainment

education. Further consideration is required of the actual educational attainment of the population as a whole, which suggests that for law to be more accessible to a larger proportion of those who read (or might read it in future), the law should be designed to be readable to those whose education is limited to secondary.

4.8. Language

Users of the site were asked to identify the language they spoke best. 93.6% identified English as their primary language. 2.01% identified Spanish, while 4.36% nominated 'other language' as their primary language. In the US, the population is primarily English speaking. However in 2011, 37.6 million people in the US spoke Spanish at home (about 12.9% of the population as a whole). Of these 25.9% self-identified as not speaking English well. The usage of the site by language may suggest lower access to law for the US population which is primarily Spanish speaking. This issue also has geographical implications as the Spanish speaking population is not uniformly distributed throughout the US, but is particularly concentrated in western and southern US states.(Ryan, 2013)

4.9. How does reading difficulty vary by demographic groups?

Although three readability datasets were collected, cloze results are most reliable as a measure of reading difficulty when used to draw comparisons between different demographic groups. Cloze tests, in contrast

to likert and semantic differential tests, provide an objective measure of reading difficulty for a reader. Likert tests and other subjective responses are affected by user perceptions and background as well as the test stimulus which may produce unreliable results when comparing between different demographic groups.

The results below are based on raw correct score results for cloze tests broken down by demographic groups (rather than proportion of correct scores).

4.9.1. *Cloze results by persona*

Figure 11 shows average correct scores for different personas for different corpora. The 95% confidence interval of the mean is also shown.[13]

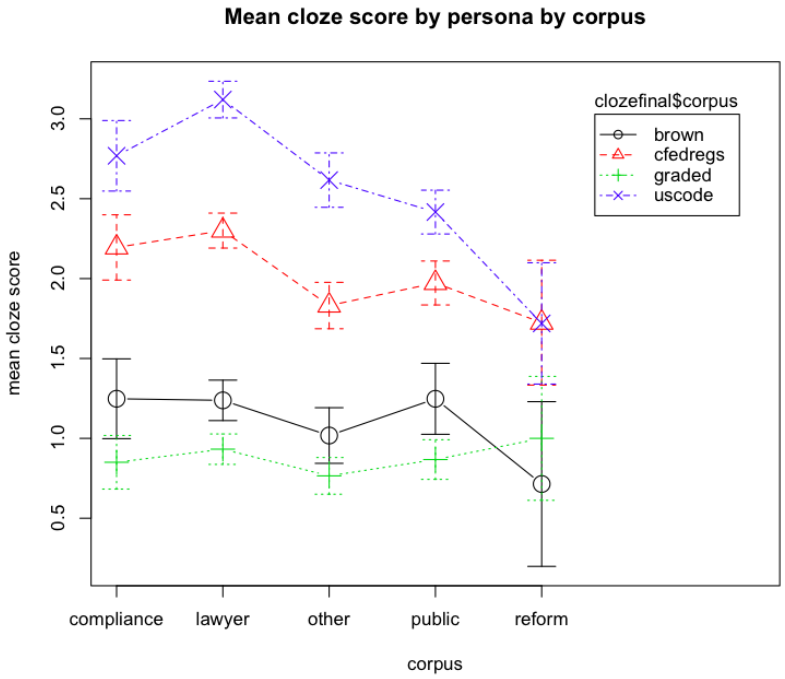

Figure 11. Mean cloze correct with 95% confidence intervals by persona by corpus

By visual inspection we can see that mean results for lawyers and the public and 'other' groups are significantly different for sentences from the US code and the Code of Federal Regulations. It is also notable that the demographic differences for the graded sentences and sentences

[13] Note that Figure 11 cannot be read to compare the corpora against each other. This is because each corpus has a different distribution of sentence lengths and the cloze scores are dependent on sentence length (e.g. there is a higher proportion of short sentences in the graded corpus than in the US Code).

from the Brown corpus are not significant. Figure 12 shows confidence intervals by persona for the US Code. This last diagram allows us to conclude that lawyers do better than all groups, including the reform or democratic element. The reform group showed the greatest difference in means with the legal group.

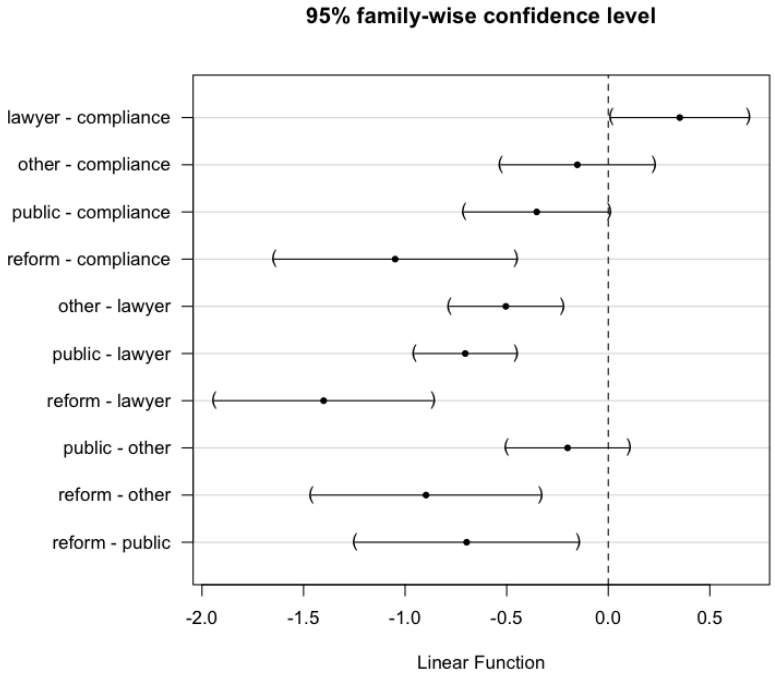

Figure 12. Mean cloze 95% confidence intervals by persona for the US Code

A one-way ANOVA test was also carried out to test for significance of differences on the mean score for the US Code for different personas finding that the mean differed significantly by persona, F (4,3049) = 23.47 , $p =< 2e - 16$ with effect size 0.023 (i.e. small).

To use the cloze tests to measure language difficulty as per the standard cloze readability methods developed by Bormuth and others (see Section 2.8 above) we need to calculate the proportion of correct responses for cloze tests of the same length. This can be done by taking a subset of data - e.g. the data in which all tests have ten gaps. Such a filtered dataset was prepared. Also sentences from the Brown corpus were removed from this subset to provide a set of cloze tests solely on sentences from the two legal corpora.

This dataset consisted of 2556 cloze test responses. Lawyers achieved average cloze proportal score of 0.42 while members of the public, compliance, and democratic groups achieved 0.35, 0.39 and 0.26, respec-

tively. There is a significant difference of means between lawyers and other groups (at $p < 0.001$) in all cases except for the difference between lawyers and the compliance group (which was not significant). Interestingly, those involved in the democratic process achieve the lowest proportional results. In Section 2.8, it was noted that results between 0.35 and 0.49 indicate the reader needs assistance to comprehend the material. Results lower than 0.35 indicate the reader is frustrated by the material. Our results, which point to legal materials being very hard to incomprehensible for many audiences, are consistent with studies described in Section 2.3 which discuss the readability or otherwise of legislation.

4.9.2. *Cloze results by other demographics*

For reasons of space, differences in readability difficulty for other demographic groups are not discussed. However the following is interesting to note. On average, women obtained a higher average cloze result than men. This was largely a result of women performing significantly better than men on cloze tests on the graded corpus.

4.10. Subjectivity, likert results and semantic differentials

Although useful for between sentence comparisons, likert results are subject to a number of issues when used for comparison across different demographic groups. For example the desire to agree with the questioner varies between cultural and other groups. In our study this effect can be very clearly seen as between Spanish speaking respondents and other language groups. Figure 13 shows level of 'agreement' by likert question. The x-axis shows the question type (i.e. whether the test sentence was easy, hard, very easy or very hard to read). The y-axis shows mean response for each question type by corpus. A lower mean indicates a higher level of average agreement. The graph shows that Spanish speakers are more likely to agree with the questioner, irrespective of question asked. Accordingly comparisons between demographic groups need to be approached with caution.

Semantic differentials are also potentially affected by the subjectivity of individual responses. For example average results for semantic differentials for gender show that women rate sentences as less readable than men, yet we saw above that in terms of mean cloze results, women scored higher than men.

Note that neither likert nor semantic differentials showed demographic differentiations in readability difficulty between demographic groups that are evident with the cloze test results. This 'subjective'

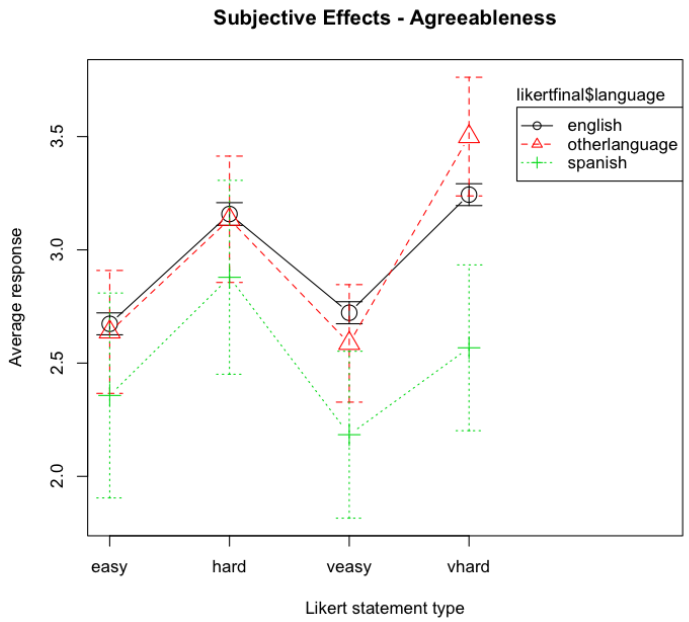

Figure 13. Subjectivity effects between demographic groups

result is interesting to compare with the study by Kandula et al. and de Clercq et al., which compared expert and non-expert evaluation of sentences. Both studies found that experts and non-experts tended to have a high level of agreement when asked to give their subjective judgement as to difficulty.(Kandula and Zeng-Treitler, 2008; De Clercq et al., 2013)

5. Measuring the Difficulty of Sentences

In this section, we turn to the question of measuring the difficulty of sentences. The data collection phase provided three datasets that could potentially contribute to the development of a ranking of sentences by difficulty (the likert, cloze and semantic differential datasets). We systematically examine each of the datasets. We also consider how the results from each set can be combined into a final measure.

5.1. Likert results

The likert dataset consists in reality of four sub-sets of data depending on the question that the user was asked.

The likert dataset was the largest, as it was most often responded to by research participants. On average, 17.86 responses were provided for each sentence, with a standard deviation of 4.39. The distribution of the number of responses by sentence was approximately normal.

Figure 14 shows the distribution of degree of "agreement" depending on the question the participant was asked. The x-axis represents degree of agreement, with 1 being 'strongly agree' and 5 being 'strongly disagree'. The y-axis peaks show proportion of responses by question type.[14] Broadly, the distribution of responses between easy questions and hard questions mirror each other. This is consistent with our intuitions as to how meaningful responses should be distributed.

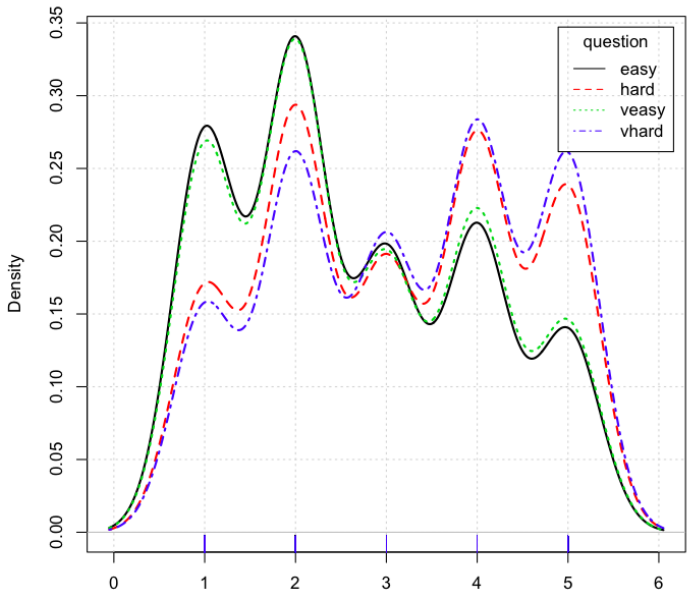

Figure 14. Density distribution of responses by question

It is also helpful to visualize average level of agreement by question and corpus, together with their 95% confidence intervals. Figure 15 provides basic validation that user responses can be used to distinguish sentences by level of reading difficulty. i.e. the averages for corpora are consistent with our expectations for the difficulty of each corpus. It can also be seen from the confidence intervals of the means that in all

[14] This graph was produced using the R statistical package. For visualizing the bandwidth (width of waves) has been artificially increased to aid visualisation. Note that 'neutral' responses and 'not sure/not applicable' responses have been combined in our analysis.

cases there is a significant difference of means between the legal and non-legal corpora, except for the "very easy" question, which in the case of the difference between the Brown Corpus and code of federal regulations did not show a significant difference at the 95% confidence level.

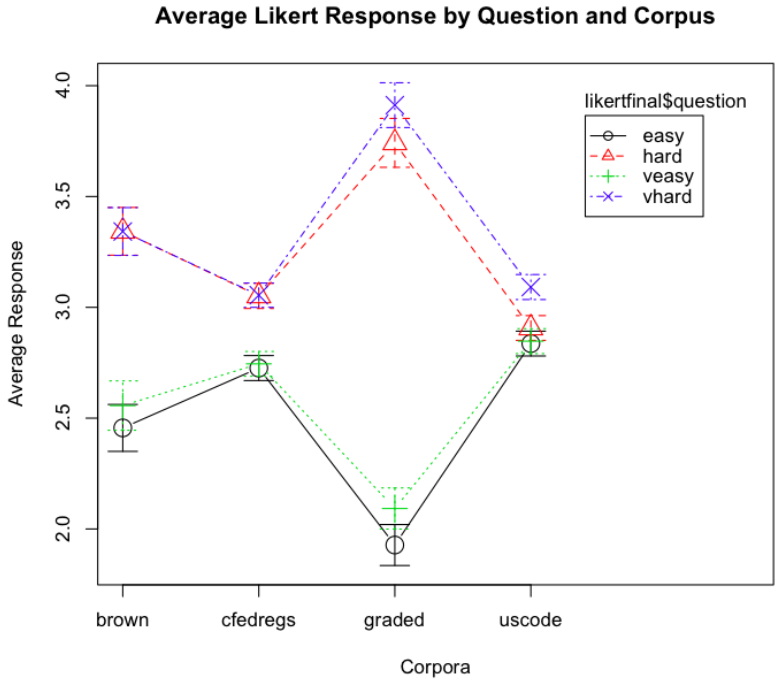

Figure 15. "Mean" likert response by question and corpus with 95% confidence intervals of the mean

To assign a composite likert measure of reading difficulty to each sentence by combining the results of the various responses, principal components analysis was applied. (See discussion above in Section 2.10) The input variables were the proportion of responses for each category in the likert test (i.e. the 20 categories (5 possible responses x 4 question types). After extracting the principal components we examined a scree plot for the data. This identified the first principal component (i.e. the component before the bend in the scree plot) as sufficient to represent the variance in the data. (See figure 16)

As a sanity check, this first principal component was compared with an aggregate measure derived by calculating the proportion of 'votes' from users indicating a sentence was 'hard' less the proportion of 'votes' that a sentence was 'easy'. This was done by first recoding

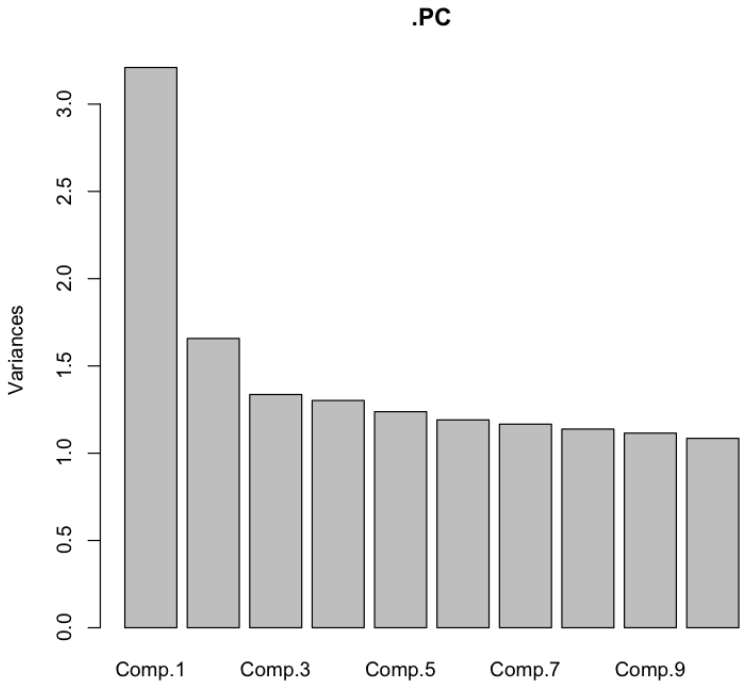

Figure 16. Scree plot of principal components extracted from likert data

and binning each response into a "hard", "easy" or "neither" vote and then calculating the proportion of votes cast for a sentence in each bin calculated. For example 'strong agreement' that a sentence is hard is classified as a 'hard' vote; and 'strong disagreement' that a vote is easy is also classified as a 'hard' vote. This 'hard-easy' variable is highly correlated with the proportion of hard votes (at 0.96) and the proportion of easy votes (at -0.97). It is also correlated at 0.95 level with the first principal component described above. Notably, the second and third principal components had low correlation with any of these measures (the highest correlation being 0.22). Principal component 3 did however correlate at -0.72 with the proportion of 'neither' votes (i.e. votes where a participant did not indicate that the sentence was either hard or easy).

We further explored the first principal component by breaking down the data into the four corpora. Figure 17 illustrates the distribution of the first principal component for sentences for each of the four corpora.

It will be evident that the metric is broadly normal and that the four corpora have different means.

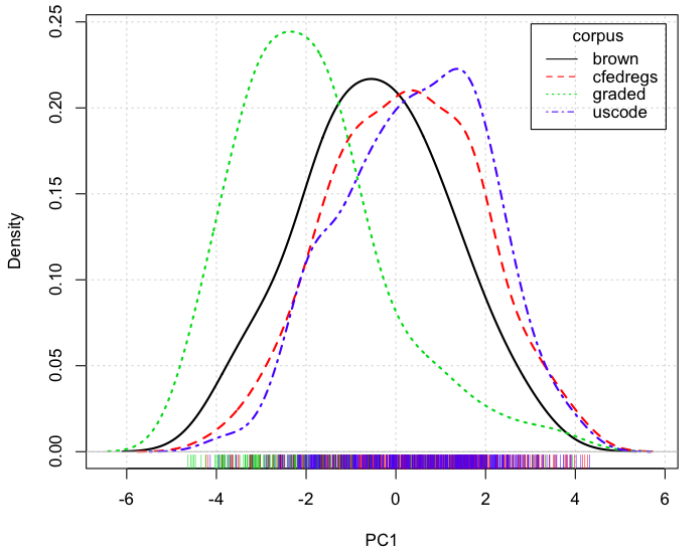

Figure 17. Density distribution of first principal component from likert results by corpus

ANOVA testing on differences between the mean results of the first principal component was carried out, as well as pairwise mean comparison. ANOVA returned a significant difference as did a comparison between all corpora (at $p < 0.001$) except between the US Code and the Code of Federal Regulations.

It is also worth noting the overlapping distributions of the corpora. Language is not sharply delineated into 'hard' and 'easy' categories: rather each sentence falls on a continuum of difficulty. This is relevant to the task of sentence classification used in machine learning which by its nature requires data to be assigned to categories. In reality reading difficulty does not come in neat separate packages that are easily detected. Most sentences are found close to a mean readability value.

5.2. Cloze results

We now turn to an analysis of the cloze results, similarly for ranking sentences by language difficulty and assigning a difficulty level to each sentence.

Sentence length strongly affects the cloze results for each corpus. This is due to the number of missing words to be guessed being dependent on sentence length. If a sentence is 50 words or less in length, the number of words to be guessed varies between one and ten. Given this, a score of '1' for a short sentence is not equivalent to a score of '1' for a longer sentence. To address this issue, results have been scaled to produce an adjusted score using the following formula:

$$\text{adjusted score} = \frac{(score + 1)}{gaps}$$

Adding 1 to the score ensures that '0' results are also scaled depending on the number of gaps in the sentence. Figure 18 compares adjustment of the score for a simple proportion (i.e. score/gaps) as compared to the formula above. A simple proportion does not produce a reasonable scaling, whereas the selected formula smoothly adjusts results by number of gaps. If the score is adjusted by a number greater than one, more extreme scaling is obtained. The optimal level of scaling may be different to that chosen, but in the absence of an external metric, the scaling to be chosen in our study is essentially an arbitrary choice. The adjustment however improves on raw scores or simple proportional adjustment.

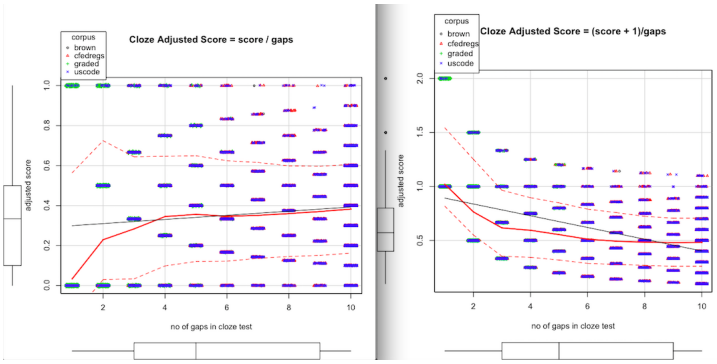

Figure 18. Comparison of cloze score adjustment schemes

An optimal scaling model merits further investigation but limitations of time made made this impractical.

The resulting ordering of sentences is different to the ordering established by the likert first principal component but is moderately correlated with it at -0.54. This level of correlation is similar to the correlation between the likert first principal component and the results discussed below from the semantic differentials (which are not affected by scaling issues).

Extending the principal components analysis described above to include the adjusted cloze results as an additional input variable produces a new first principal component, which correlates with the likert first principal component result at 0.98.

5.3. Semantic differential results

Semantic differentials can be used to derive a measure of user experience, which can also be broken down into a more nuanced set of characteristics. Semantic differential results were collected on responses to ten different semantic opposites. The semantic opposites were primarily chosen to capture user experience of using law, but also sought to explore the three dimensions which were identified by Osgood in his studies of semantic differentials: evaluative (good-bad); power (strong-weak): activity (active-inactive).

Six of the semantic differentials addressed concerns central to usability or user experience or readability of law. Two others addressed characteristics that users might associate with law: fairness-unfairness and attraction-repulsion. Both are evaluative, but evaluate law against notions of equity or emotional response to the content of the law. These two characteristics were chosen to explore whether individuals thoughts/feelings about the content of law affects their assessment of its readability. A list of the semantic differentials used is provided below. A summary term is provided in brackets and the semantic differential is asterisked if concerned with usability/user experience/readability characteristics. Although the original scale was between -3 and +3, the scale was adjusted to range between 1-7. Also, where necessary, scales were flipped so that a higher result means increasing strength in the characteristic. e.g. a readability score of 1 indicates less readability than a readability score of 7. This recoding was to assist in analysing and communicating results.

attractive-repellant (attractiveness)

clear-obscure (clarity)*

fair-unfair (fairness)

familiar-strange (familiarity)*

helpful-unhelpful (helpfulness)*

interesting-dull (interest)

severe-lenient (leniency)

readable-unreadable (readability)*

complex-simple (simplicity)*

usable-unusable (usability)*

Pearson's correlations were calculated for both raw semantic differential scores and mean scores for sentences. The level of correlation follows broadly the same pattern, with higher correlations found for the averaged scores (which is consistent with averaging out individual response variance). Table III shows correlations for the six semantic differentials associated with user experience and also includes correlation with the average adjusted cloze score and the likert first principal component. As will be evident, correlations across the table are moderate to high for most given characteristics, though the cloze average correlation was low for clarity, familiarity, helpfulness and readability and moderate with likert, readability and simplicity.

The pattern of correlation suggests that user experience characteristics including readability are not perfectly aligned although these characteristics have some degree of correlation. They also suggest that different testing methods will evoke different patterns of responses from users. Notably, the subjective measures (semantic differentials and likert tests) align to a greater degree than does the objective measure (cloze testing), although the likert and cloze results correlate moderately as between themselves. Among the semantic differentials we used, clarity, simplicity, readability and familiarity showed the highest correlation with the likert and cloze results. Also readability and clarity were highly correlated (0.83) and helpfulness and usability were also highly correlated (0.83). Simplicity had low to moderate correlation with helpfulness and usability (0.35 and 0.37, respectively). The differences in correlation indicate the more detailed description of user experience that semantic differentials can provide.

It is also of interest to examine distributions of responses for factors not associated with user experience, in this case looking at frequency of raw scores submitted by users. The diagonal on figure 19 shows this distribution. In most cases, users did not regard factors such as leniency or fairness as being relevant to assessing a test sentence. This may be compared to the quite different distribution for clarity, which did evoke mainly positive or negative assessments of clarity.

Table III. Pearson's R correlation for adjusted average cloze score, likert first principal component and six semantic differentials most related to user experience.

	clozemean	likertPC1	clarity	familiarity	helpfulness	readability	simplicity	usability
clozemean	1.00	-0.55	0.34	0.33	0.13	0.46	0.53	0.13
likertPC1	-0.55	1.00	-0.59	-0.53	-0.38	-0.65	-0.62	-0.40
clarity	0.34	-0.59	1.00	0.70	0.74	0.83	0.62	0.74
familiarity	0.32	-0.53	0.70	1.00	0.62	0.69	0.51	0.65
helpfulness	0.13	-0.38	0.74	0.62	1.00	0.64	0.35	0.83
readability	0.46	-0.65	0.83	0.69	0.64	1.00	0.68	0.65
simplicity	0.53	-0.62	0.62	0.51	0.35	0.68	1.00	0.37
usability	0.13	-0.40	0.74	0.65	0.83	0.65	0.37	1.00

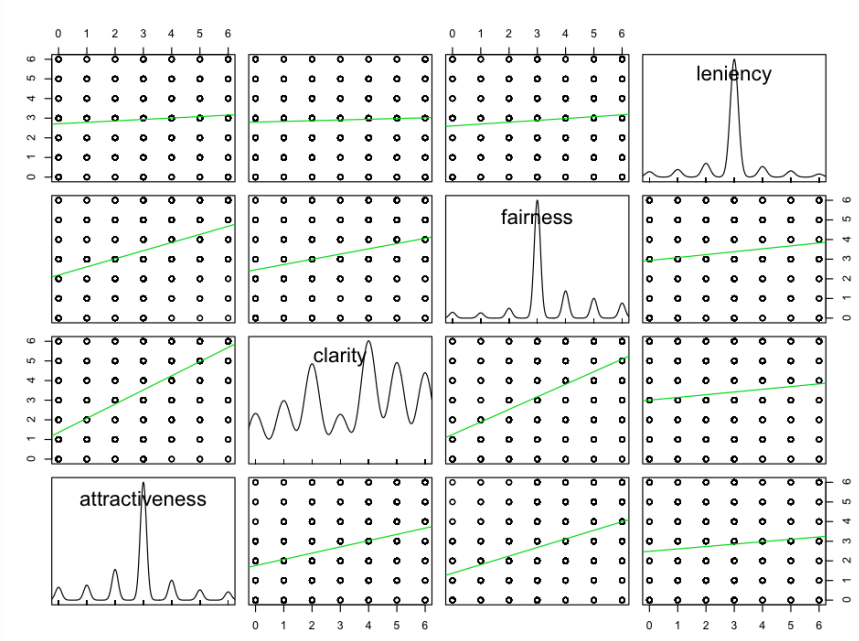

Figure 19. Density distribution of non-user experience characteristics

5.4. A Total Composite Readability Measure - Multivariate Analysis

In this section, we develop a total composite measure of readability. As with the individual measures, this measure can be developed by em-

ploying principal components analysis to extract the primary dimension of variance.

Table IV show the correlation between various composite measures of sentence difficulty. Clozemean1 is the mean adjusted cloze result described above. Semdiff4PC1 is the first principal component of semantic differential results using only the four semantic differentials which were most highly correlated with the cloze and likert results for this input: i.e. clarity, familiarity, readability and simplicity. Notably, these are semantically the most similar to the concept of easy/hard to read used in likert testing (usability and helpfulness being the other relevant terms which were less correlated). LikertPC1 is the first principal component of the likert results. CompositePC1 is the first principal component of the combined results from all tests. This last measure was derived from all 20 likert variables, the four semantic differential variables and the cloze mean adjusted score. The resulting composite measure is highly correlated with the principal component for semantic differentials and likert results, and moderately correlated with cloze mean results. The scree plot for this composite measure is also shown below (Figure 20), and as in the case of the Likert results, the first principal component is the sole component that satisfies the scree plot test.

Table IV. Correlation between different measures of 'readability'

	clozemean1	compositePC1	likertPC1	semdiff4PC1
clozemean1	1.00	-0.64	-0.55	-0.47
compositePC1	-0.64	1.00	**0.92**	**0.90**
likertPC1	-0.55	0.92	1.00	0.68
semdiff4PC1	-0.47	0.90	0.68	1.00

The final composite measure can be used to generate an ordering of sentences which can be used to assign sentences to an 'easy' and 'hard' difficulty classifications. A 50% dividing line was used, thus 'easy' simply means the easiest half of the sentences, and 'hard' the hardest half. The assigned classifications can then be used as an input to machine learning.

It is worth commenting at this point that throughout the data, high variance was encountered in individual responses. How hard a user perceived or experienced a sentence to be, varied widely for both subjective and objective tests. The ability to derive reliable comparative measures of sentence difficulty therefore depends on being able to collect a sufficient number of assessments from users for each individual sentence.

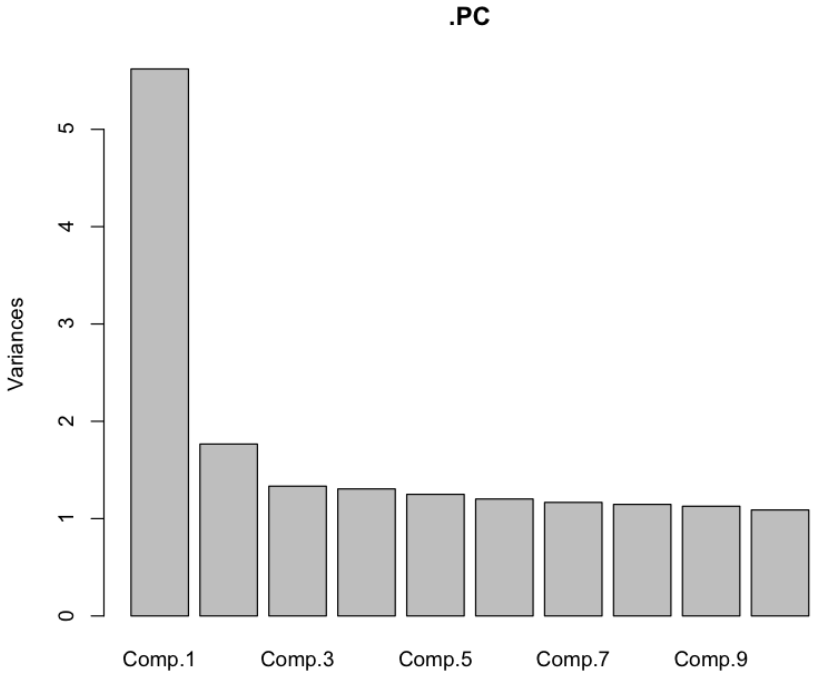

Figure 20. Composite principal components scree plot

Our results suggest that at least 15-20 separate user evaluations need to be collected for each test sentence.

6. Machine Learning

Collecting user evaluations of sentences is time consuming and difficult. It requires the availability of online infrastructure and access to audience. It calls on the time of users who are asked to provide evaluations. As this study has demonstrated, it is feasible to collect such user evaluations and this can provide valuable insight into the readability of legal language. Ideally however, we would wish to be able to predict the reading difficulty of a sentence without having to conduct surveys.

As has been noted, we divided the sentences into two equally sized "easy" or "hard" classes, depending on the sentence ranking according to the composite measures described in Section 5.4. To investigate

application of machine learning we extracted natural language features from the sentences including:

(a) sentence length;

(b) average word length;

(c) type to token ratio (i.e. ratio of unique words to total words);

(d) common readability metrics;

(e) proportion of verbal phrase chunks; and

(f) proportional distribution of different parts of speech.

Tests were carried out on two datasets: a dataset of all four corpora and a dataset of the sentences from just the two legal corpora. After a number of trials to investigate which machine learning algorithm achieved the highest level of accuracy for this task; a support vector machine (SVM) was chosen for the learning task.[15] All tests were validated using 10-fold cross validation.

6.1. Results for four corpora dataset

An accuracy (F-measure) of was 72.7% was achieved. The overall result is less than the results reported at sentence level by Dell'Orletta et al which are discussed above in Section 2.6, who report an accuracy of 78% on the task of classifying sentences.

We also investigated the effect of removal (ablation) of particular features on prediction accuracy. In particular, we explored the contribution of traditional readability metrics to accuracy of machine learning.

6.1.1. *Effect of Ablation on Machine Learning Accuracy for Four Corpora Dataset*

For the whole dataset accuracy was reduced to 67.1% if only readability metrics were used as input. This was about the same accuracy as was achieved using just sentence length and average word length (67.3%). This is not surprising as readability metrics depend heavily on these

[15] That is, using the SMO package which is the support vector machine implementation in the Weka software. All reported results are for an SVM. The intuition behind an SVM is that (in a two dimensional case) the algorithm seeks to find the dividing line that maximises the distance of data points from the dividing line. In a case with many input features (which is usual for machine learning), the 'dividing line' is actually a hyperplane and each input feature is a dimension of a multidimensional space.

Table V. Precision, Recall and F-measure Support Vector Machine (SMO) on Four Corpora Dataset

Class	Precision	Recall	F-Measure
easy	0.737	0.705	0.72
hard	0.718	0.749	0.733
Weighted Avg.	0.727	0.727	0.727

two features. Using all features *except* readability metrics resulted in a prediction accuracy of 72.5%, i.e. virtually the same as including readability metrics with other as input features. Readability metrics can be concluded to be useless in predicting classification from the composite measure we used.

6.1.2. *Precision, Recall and F-measure Support Vector Machine (SMO) on Legal Corpora Dataset*

On purely legal sentences accuracy was 70.5%, i.e. a little less than for the four corpora dataset. Again this matches expectations as the graded sentences are virtually all in the easy dataset.

Table VI. Precision, Recall and F-measure Support Vector Machine (SMO) on Legal Corpora Dataset

Class	Precision	Recall	F-Measure
easy	0.691	0.571	0.625
hard	0.713	0.807	0.757
Weighted Avg.	0.703	0.705	0.7

6.1.3. *Effect of Ablation on Machine Learning Accuracy for Legal Corpora*

For the legal dataset accuracy was reduced to 60.2% if only readability metrics were used as input. Using all features except readability metrics resulted in an accuracy of 70.5%. For the legal dataset accuracy was 56% using just sentence length and average word length (i.e. machine learning essentially failed). Using just average word length, sentence length and type to token ratio achieved an accuracy of 66%. While using just phrase proportions and parts of speech proportions attained an accuracy of 67.8%

To further investigate the relationship between language difficulty and readability metrics we examined correlation between readability metrics and the composite difficulty measure. The SMOG index was most highly correlated at 0.33, which was about the same as correlation for sentence word length. This was however exceed by the type to token ratio at -0.42. In other words, it is more effective to count the ratio of unique words to total words as a measure of language difficulty of legal sentences than to rely on readability metrics at sentence level. Further, both for the four corpora dataset and the legal corpora dataset more accurate predictions can be obtained by a machine learnt model, than by using traditional readability metrics.

6.2. Discussion of machine learning results

The results of machine learning show the feasibility of improving accuracy of readability prediction over traditional readability metrics (i.e. 70.5 versus 60.2). This result is consistent with findings reported in the research literature. As far as we are aware this result has not been applied previously to legislative language (which is of course a unique form of English). We were able to show that accuracy can be increased on legal sentences by about 10% over use of traditional readability metrics alone. The overall level of accuracy of 70.5% is lower than that reported by other researchers on sentence level classification. It may be possible to increase the level of accuracy by extracting more complex natural language features. Also, it is likely that increasing the amount of data on which learning can be carried out would also increase accuracy. The rate at which we could generate human labelled data was as limiting factor in our study.

7. Conclusions

7.1. Applying citizen science to readability of legislative texts

The research reported in this paper demonstrates the feasibility of using citizen science (in the form of online crowdsourced data collection) to create a corpus of labelled data for input to machine learning for predicting the readability of legal rules. It is possible to rank a given set of legal sentences by reading difficulty using responses submitted by users. However, the time required to collect the necessary data is non-trivial, even on a large sites such as LII. Data in our study was collected over a three month period. Even after three months, the sentences tested represents a tiny proportion of the legal language

found in the United States Code and Code of Federal Regulations. Extending this research to a larger dataset would potentially require orders of magnitude longer. This is far from a fatal barrier, and does give insight as to what may be required to collect sufficient data to better predict the readability of legal language. However, considering the time horizons that are sometimes necessary for research in other fields (for example longitudinal health studies, or multiyear collection of astronomical data), it is well within the bounds of realistic research. An observation that bears on this conclusion is that participant responses were maintained throughout the period of collection, suggesting that it is feasible to collect data over long periods, without the rate at which data is collected reducing over time. A novel (or at least unusual) aspect of our citizen science project was that citizen scientists assisted in collecting information about themselves as well as about the 'data'.

7.2. Demographic insights

The demographic results described above are also of interest. We can begin to reach conclusions about who reads the law and why they do so. Legal professionals (including law students) were a minority of those who participated as citizen scientists in our research. Determining whether this is true of the online audience for legislation generally, requires further research, including on other online sites. Nonetheless, a substantial audience for online legislative materials are non-lawyers. Non-lawyers find legislative materials harder to access that the legally trained. If we wish to communicate effectively with this substantial audience, we need to re-examine how the law is written. Women, those not having tertiary education and those for whom Spanish is the primary language were under-represented among participants. Gender, education and language aspects of access merit further investigation. Our study is consistent with the findings of other researchers that legislative language is harder for those without legal training. It also suggests that legislative language is hard for all audiences, including the legally trained.

7.3. Machine learning

Our work on machine learning reports initial application of machine learning to the readability of legislative materials. We have demonstrated that traditional readability metrics can be improved on, for high resolution (i.e. sentence level) automatic classification of legal sentences into a binary easy vs. hard classes. The level of accuracy attained is moderate and would require further improvement to provide a reasonably usable automated detection system. We are planning to publish a

second paper that extends the machine learning results reported here, including investigating how accuracy may be increased by increasing the number of input features and further exploring whether increasing the available data may assist in improving accuracy.

7.4. Methods of Measuring Readability Using Crowdsourced Data

Part of this paper is devoted to describing the methods we used to collect crowdsourced assessments of readability. A considerable portion also describes the methods used to convert crowdsourced assessments into measures of sentence readability. In part this goes to reproducibility of the research. However, it also seemed useful to us to describe our methods at some length, as there is no 'standard' method for carrying out crowdsourced readability research. It is useful for the research literature to provide descriptions of this kind. Further, we have no doubt that the methods reported in this study can be improved on.

Of the three tests that we used, likert tests proved to be the most effective in attracting participation and in ensuring data was usable. Semantic differentials provided a more nuanced characterisation of the sentences being tested. However, the rate of response for semantic differentials was much lower than likert tests and the occurrence of unreliable data much higher. Cloze tests, unlike the previous two tests, had the advantage of providing an objective measure and proved to be the only useful test for distinguishing readability for different demographic groups. However, analysis of cloze results was complex and like semantic differentials they attracted a lower response rate. Again issues of data reliability reduced the usable data. Also given wide differences in sentence length, cloze tests were not ideally suited to sentence level assessment. A limiting factor that emerged in the study, for all three methods, was the rate at which assessments could be collected. Methods which reduce the necessity for replication may significantly increase the rate of data collection.

7.5. How readers read the law online

We may be all equal before the law but the law is not equally of interest to its readers. In fact, the frequency with which a particular piece of law is read follows a power law distribution. This is an important insight. If we are concerned with improving reader experience, attention to that part of the law which is most read, provides exponentially greater return and requires fractional effort as compared to seeking to improve the law book as a whole. Further, if law is not being read, we may ask the question: how important is it for that law to remain in the

law book? At least, in terms of organising legal documents, those parts which are most read, might be usefully reorganised in ways that make them more accessible to readers.

7.6. Some broader implications

The data repositories and online publishing platforms which sites such as Cornell LII maintain, can perhaps be thought of as potentially playing enhanced roles in improving access to law. Such sites have achieved access to law in terms of ensuring that citizens are able to find and access law online. The fact of availability does not, however, necessarily equate to "access" in all its senses. Addressing the readability of legislation by applying online technologies is a natural extension of the work already carried out by the Free Access to Law Movement.

Online legal publishing platforms are also potentially sites for the ongoing collection of data which illuminates how users interact online with legal language. They are not simply collections of text or collections of data, they are a focus of a dynamic and ongoing interaction between human beings and the laws that govern them. We can perhaps trace the outlines of a paradigm in which the publication of law online - already moving from being conceived as static document to data repositories - is reconceptualised even further as an online platform capturing a multiplicity of points of human-legal interaction with the potential to tell us a great deal about the social dimensions of law. Or, in other words, online law is part of a social network in which both human beings and legal rules (communicated by other human agents) are nodes. The insights that we may derive from a study of these interactions could over time be applied to improve legal language – addressing an as yet unmet dimension of making law accessible to all who would like to have that access. To extrapolate from the words of the UK Parliamentary Counsel: when citizens find the law, they should able to read it. Other applications outside the readability field may also exist.

8. Future Work

We are interested in extending the work reported here into the following areas of research:

(a) extending citizen scientist participation in other aspects of readability research (for example project design);

(b) investigating other means of collecting readability assessments of legal language online, for example A-B testing, a simplified form

of likert or approaches that calibrate between different testing approaches;

(c) further investigating the demographic aspects of access to law online, particularly gender, education and language; and

(d) extending the preliminary machine learning results reported in this paper.

References

Eloise Abrahams. Efficacy of plain language drafting in labour legislation. Master's thesis, 2003.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics, 2010.

Gregory Asmolov. Crowdsourcing as an activity system: Online platforms as mediating artifacts. In *Sintelnet WG5 Workshop on Crowd Intelligence: Foundations, Methods and Practices*, 2014.

Robert W Benson. End of legalese: The game is over. *NYU Rev. L. & Soc. Change*, 13:519, 1984.

J. Bentham. Nomography, or the art of inditing laws. *The Works of Jeremy Bentham (ed. J. Bowring)*, 3:231 et seq, 1843.

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

J.R. Bormuth. *Cloze readability procedure*. University of California Los Angeles, 1967.

F. Bowers. Victorian reforms in legislative drafting. *Tijdschrift voor Rechtsgeschiedenis*, 48:329, 1980.

Daren C Brabham. Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: the international journal of research into new media technologies*, 14(1):75–90, 2008.

Dennis L Clason and Thomas J Dormody. Analyzing data measured by individual likert-type items. *Journal of Agricultural Education*, 35:4, 1994.

Kevyn Collins-Thompson. Computational assessment of text readability: A survey of past, present, and future research. *working draft*, July 2014.

Kevyn Collins-Thompson and James P Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200, 2004.

Anna B Costello and Jason W Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7):2, 2005.

Mick P Couper, Roger Tourangeau, Frederick G Conrad, and Eleanor Singer. Evaluating the effectiveness of visual analog scales a web experiment. *Social Science Computer Review*, 24(2):227–245, 2006.

Michael Curtotti and Eric McCreath. Enhancing the visualization of law. In *Law via the Internet Twentieth Anniversary Conference, Cornell University*, 2012.

Michael Curtotti and Eric McCreath. A right to access implies a right to know: An open online platform for research on the readability of law. *Journal of Open Access to Law*, 1(1), 2013.

O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken. Using the crowd for readability prediction. *Natural Language Engineering*, 2013.

Joost CF de Winter and Dimitra Dodou. Five-point likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11):1–12, 2010.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83. Association for Computational Linguistics, 2011.

Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4): 86–96, 2011.

W.H. DuBay. The principles of readability. *Impact Information*, pages 1–76, 2004.

M. Evans and R.I. Jack. *Sources of English Legal and Constitutional History*. Butterworths, 1984.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

Carolina Ferrari, Tomas Garcia & Short. Legibility and readability on the world wide web, 2002. URL `http://bigital.com/english/files/2008/04/web_legibility_readability.pdf`.

Frank J Floyd and Keith F Widaman. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, 7(3):286, 1995.

John Fox. The R Commander: A basic statistics graphical user interface to R. *Journal of Statistical Software*, 14(9):1–42, 2005. URL `http://www.jstatsoft.org/v14/i09`.

W. N. Francis and H. Kucera. A Standard Corpus of Present-Day Edited American. Revised 1971, Revised and Amplified 1979. Department of Linguistics, Brown University Providence, Rhode Island, USA, 1964. URL `http://www.hit.uib.no/icame/brown/bcm.html`.

Michael S Friman. Plain english statutes-long overdue or underdone. *Loy. Consumer L. Rep.*, 7:103, 1994.

Edward Fry. A readability formula for short passages. *Journal of Reading*, 33(8):594–597, May 1990.

Ron Garland. A comparison of three forms of the semantic differential. *Marketing Bulletin*, 1(1):19–24, 1990.

GLPi and V. Smolenka. A Report on the Results of Usability Testing Research on Plain Language Draft Sections of the Employment Insurance Act, 2000. URL `http://www.davidberman.com/wp-content/uploads/glpi-english.pdf`.

Elizabeth M Grieco, Yesenia D Acosta, G Patricia de la Cruz, Christine Gambino, Thomas Gryn, Luke J Larsen, Edward N Trevelyan, and Nathan P Walters. The foreign-born population in the United States: 2010. *American Community Survey Reports*, 19:1–22, 2012.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software. *SIGKDD Explorations*, 11(1), 2009.

Richard H Hall and Patrick Hanna. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23 (3):183–195, 2004.

Eric Hand. Citizen science: People power. *Nature International Weekly Journal of Science*, 466:685–687, 2010.

Wolfgang Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Published online, 2003.

J. Harrison and M. McLaren. A plain language study: Do New Zealand consumers get a "fair go" with regard to accessible consumer legislation. *Issues in Writing*, 9:139–184, 1999.

Richard M Heiberger and Naomi B Robbins. Design of diverging stacked bar charts for likert scales and other applications. *Journal of Statistical Software submitted*, pages 1–36, 2013.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79. Association for Computational Linguistics, 2008.

P. Heydari and A.M. Riazi. Readability of texts: Human evaluation versus computer index. *Mediterranean Journal of Social Sciences*, 3 (1):177–190, 2012.

Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

Simon James and Ian Wallschutzky. Tax law improvement in Australia and the UK: the need for a strategy for simplification. *Fiscal Studies*, 18(4):445–460, 1997.

Frances Johnson. Using semantic differentials for an evaluative view of the search engine as an interactive system. In *EuroHCIR*, pages 7–10, 2012.

Sasikiran Kandula and Qing Zeng-Treitler. Creating a gold standard for the readability measurement of health texts. In *AMIA Annual Symposium Proceedings*, volume 2008, page 353. American Medical Informatics Association, 2008.

Tapas Kanungo and David Orr. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International*

Conference on Web Search and Data Mining, volume WSDM '09, February 9-12, 2009, Barcelona, Spain., pages 202–211. ACM, 2009.

Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 546–554. Association for Computational Linguistics, 2010.

David Kauchak, Obay Mouradi, Christopher Pentoney, and Gondy Leroy. Text simplification tools: Using machine learning to discover features that identify difficult text. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 2616–2625. IEEE, 2014.

J. Kimble. Answering the critics of plain language. *Scribes J. Leg. Writing*, 5:51, 1994.

G.R. Klare. Readable computer documentation. *ACM Journal of Computer Documentation (JCD)*, 24(3):148–168, 2000.

S. Krongold. Writing laws: Making them easier to understand. *Ottawa L. Rev.*, 24:495, 1992.

Bruce Lewenstein. What does citizen science accomplish. 2004.

D. Melham. Clearer Commonwealth Law: Report of the Inquiry into Legislative Drafting by the Commonwealth. Technical report, House of Representatives Standing Committee on Legal and Constitutional Affairs, 1993.

David Mellinkoff. *The Language of the Law*. Little, Brown and Company, 1963.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics, 2010.

Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

PCO NZ. Presentation of New Zealand Statute Law: Issues Paper 2. Technical Report 2, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office, 2007.

PCO NZ. Presentation of New Zealand Statute Law. Technical Report 104, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office, 2008.

Law Reform Commission of Victoria. Access to the Law - the structure and format of legislation. Technical Report 33, Law Reform Commission of Victoria, 1990.

OLR. Inland Revenue Evaluation of the Capital Allowances Act 2001 rewrite, Opinion Leader Research. Technical report, UK Inland Revenue, 2003.

OPC-Australia. Plain English. Technical report, Australian Commonwealth Office of Parliamentary Counsel, 2003.

OPC-UK. When Laws Become Too Complex: A Review into the Causes of Complex Legislation. Technical report, United Kingdom Office of Parliamentary Counsel, 2013.

N. Pettigrew, S. Hall, and D. Craig. The Income Tax (Earnings and Pensions) Act - Post-Implementation Review, Final Report MORI, 2006.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics, 2008.

Marta Poblet, Esteban García-Cuesta, and Pompeu Casanovas. IT enabled crowds: Leveraging the geomobile revolution for disaster management. In *Proceedings of the Sintelnet WG5 Workshop on Crowd Intelligence: Foundations, Methods and Practices*, pages 16–23, 2014.

R-Core-Team et al. R: A language and environment for statistical computing. 2012.

D Renton. The preparation of legislation - report of a committee appointed by the Lord President of Council. Technical report, Council of the UK Government, 1975.

Camille L Ryan. Language use in the United States: 2011 American community survey reports. *Washington, DC: US Census Bureau*, 2013.

Camille L Ryan and Julie Siebens. Educational attainment in the United States: 2009. *Washington, DC: US Census Bureau*, 2012.

Adrian Sawyer. Enhancing compliance through improved readability: Evidence from New Zealand's rewrite "experiment". *Recent Research on Tax Administration and Compliance*, 2010.

Sarah E Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics, 2005.

Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM, 2001.

Johan Sjöholm. *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. PhD thesis, Linköping, 2012.

D. Smith and G. Richardson. The readability of Australia's taxation laws and supplementary materials: an empirical investigation. *Fiscal Studies*, 20(3):321–349, 1999.

Edwin Tanner. Seventeen years on: Is Victorian legislation less grammatically complicated. *Monash UL Rev.*, 28:403, 2002.

Carol Tullo. Solving the challenge of the 21st century statute book. In *Law via the Internet Conference*, 2013. URL `http://www.jerseylvi2013.org/presentation/solving-the-challenge-of-the-21st-century-statute-book/`.

Arnold Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pages 521–530. ACM, 2010.

G. Wagner. Interpreting cloze scores in the assessment of text readability and reading comprehension, 1986.

Andrea Wiggins and Kevin Crowston. From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

B. Woods, G. Moscardo, T. Greenwood, et al. A critical review of readability and comprehensibility tests. *Journal of Tourism Studies*, 9(2):49–61, 1998.

Acknowledgements

We would like to thank the users of the Cornell LII site who kindly and generously contributed their time to provide crowd-sourced assessments of the difficulty of legal language. We also thank the reviewers of an earlier version of this paper for their helpful feedback.