Beyond Readability with RateMyPDF*

A Combined Rule-based and Machine Learning Approach to Improving Court Forms

Quinten Steenhuis
Legal Innovation and Technology Lab
Suffolk University Law School
Boston, Massachusetts USA
qsteenhuis@suffolk.edu

Bryce Willey
Legal Innovation and Technology Lab
Suffolk University Law School
Boston, Massachusetts USA
bwilley@suffolk.edu

David Colarusso
Legal Innovation and Technology Lab
Suffolk University Law School
Boston, Massachusetts USA
dcolarusso@suffolk.edu

ABSTRACT

In this paper, we describe RateMyPDF, a web application that helps authors measure and improve the usability of court forms. It offers a score together with automated suggestions to improve the form drawn from both traditional machine learning approaches and the general purpose GPT-3 large language model. We worked with form authors and usability experts to determine the set of features we measure and validated them by gathering a dataset of approximately 24,000 PDF forms from 46 U.S. States and the District of Columbia. Our tool and automated measures allow a form author or court tasked with improving a large library of forms to work at scale.

This paper describes the features that we find improve form usability, the results from our analysis of the large form dataset, details of the tool, and the implications of our tool on access to justice for self-represented litigants. We found that the RateMyPDF score significantly correlates to the score of expert reviewers.

While the current version of the tool allows automated analysis of Microsoft Word and PDF court forms, the findings of our research apply equally to the growing number of automated wizard-driven interactive legal applications that replace paper forms with interactive websites.

CCS CONCEPTS

•Human-centered computing~Accessibility •Human-centered computing~Accessibility~Accessibility technologies •Human-centered computing~Accessibility~Accessibility design and evaluation methods •Information systems~Information retrieval~Document representation~Content analysis and feature selection •Applied computing~Law, social and behavioral sciences~Law •Applied computing~Document management and text processing~Document capture~Document analysis

KEYWORDS

Accessibility, Law, Administrative Burden, Readability, Court Forms, Automated Analysis

ACM Reference format:

Quinten Steenhuis, Bryce Willey, and David Colarusso. 2023. Beyond Readability with RateMyPDF: A Combined Rule-based and Machine Learning Approach to Improving Court Forms. In *Proceedings of International Conference on Artificial Intelligence and Law (ICAIL'23). ACM, New York, NY, USA, 10 pages.*

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595146

1 Introduction

The legal form is the primary way that self-represented litigants interact with courts across the United States. Self-represented litigants make up the vast majority of users of the civil court system. In 2015, only 24% of cases in U.S. civil courts had representation for both parties [12]. Within a single jurisdiction, court systems may provide litigants with up to 1,500 standardized forms that address different legal rights. Legal forms require untrained litigants to read, understand, gather information, and apply legal reasoning. Difficult forms place a time and emotional burden on litigants, can make it hard for judges to understand what litigants want, and can lead to unfair outcomes in court.

In some jurisdictions, a small number of forms have been converted into what are called interactive legal applications [15] or guided interviews. These expert system-like question and answer tools can greatly improve the ease of use of forms, but they take a large amount of time and effort to create. Our lab's ongoing work has been to build tools that simplify and increase the speed of automation of forms [24], but the vast majority across the United States remain available only as Microsoft Word or PDF documents.

In our experience working with courts across the country, we have observed that court forms are often created by untrained internal staff, without specialized tools. Forms are often designed in a word processor, such as Microsoft Word, or in better funded

courts, desktop publishing tools, like Adobe InDesign. A small number of state courts have a "forms committee" tasked with gathering the input of various stakeholders in the form's design. Stakeholders may include attorneys and court clerks who practice in the area the form covers. On occasion, form authors use templates and style guides to ensure that the form requests information consistently. Almost never do the form committees we have observed include:

- Self-represented users of the court system
- Designers
- Plain language and readability experts

The most reliable way to improve a form is to conduct an observational study of real self-represented litigants completing the form and then to identify areas where litigants experience difficulty (i.e., a traditional usability test). However, usability tests alone cannot provide guidance for the creation of the first draft of a new form. In addition, usability testing can be time consuming and even modest compensation for usability test subjects may be outside of a court's budget. Expert guidelines can address an important need in the creation of easy to use court forms. The guidelines discussed in this paper will help court staff and legal providers revise forms so that they are simpler, easier to understand, and easier to fill out accurately and completely. The RateMyPDF tool extends the value of the expert guidelines by helping courts quickly identify areas for improvement in either a single form or a large group of forms in an automated way. RateMyPDF allows form authors to "work at scale."

While RateMyPDF measures features of printable court forms, many of the rules and guidance that apply to printable forms are also applicable to interactive legal applications.

2 Evaluating form difficulty

What makes a form easy or hard to fill in? We propose that the difficulty of a court form depends on a typical self-represented litigant's ability to:

- Comprehend the form prompts and instructions
- Accurately provide the requested information
- Consistently provide a complete response to the form

As well as:

- The time burden imposed on the form's user
- The psychological burden or harm imposed by requiring the user to recount traumatic events

To build the guidelines below, we looked at existing written material, including the U.K.'s guidance on writing good questions [35], relied on our own experience as constructors and usability test conductors of dozens of interactive legal applications over the past 5 years, and interviews with authors of interactive legal applications and designers of court forms from 8 legal aid programs and courts from 8 different states in the United States. In this section, we discuss the guidelines and briefly identify strategies for measuring them. In the next two sections we discuss how we benchmarked each measure by evaluating forms from 46 States and the District of Columbia and then implemented each guideline in the RateMyPDF web application.

2.1 Helping users comprehend form prompts and instructions

Within important limits, a form's prompts, labels and instructions are easier to comprehend when they are written at a lower *reading grade level*. *Reading grade level* is a common metric produced by readability instruments that measure the ease of comprehension of narrative texts. In the United States, the median reader can comprehend texts written for a grade level between 8 and 9 [7,22], which means a significant percent of the population requires a lower reading grade level to easily comprehend the text. We therefore join a long tradition by recommending that form authors target writing for forms at a 6th grade reading level [16].

The concept of measuring readability and assigning it a score became popular in 1948, which is the year that the two most used measures, Flesch-Kincaid [9] and Dale-Chall [5], were first published. Both measures are friendly to computation by hand and use simple metrics. Flesch-Kincaid [9], for example, assigns texts a "grade level" score based on the length of sentences and the number of syllables in each word. The Dale-Chall formula adds a table of the most common 3,000 words in the English language [5]. Texts that include words that do not appear on the table are scored as more difficult to read. The two formulas often reach equivalent results on similar texts.

Even when measuring their target of narrative text, readability instruments have limits [1,20]. The text that people must read in forms is quite different from the narrative text that readability instruments were first created to measure. Forms are primarily composed of a mix of instructions, labels, and prompts, often without punctuation for headings and labels. Instructions are often minimal. Labels are often a single word: "Name," "Address," and so on. How to accurately turn these fragments into "sentences" that readability instruments can analyze is not clear. Forms may use common words like "Answer" and "Complaint" in legally-specific ways. These features combine to defeat the reliability of readability measures, which use sentence length and vocabulary as a proxy for complexity in forms.

Because of these limitations, readability measures are not sufficient as a final measure of the ease of comprehending a form's instructions. We suggest using readability measures as a starting point and separately measuring the difficulty of a form's instructions with the use of a vocabulary list such as the Dale-Chall difficult word list [5].

2.2 Guiding users in providing accurate and complete responses

Forms require the reader to write responses, not simply to understand and recall information. Both the substance of the

expected answer and the input type can affect both accuracy and completeness. Court forms are high stakes. Court forms, as compared to other forms that members of the public may use to interact with corporations or government, are more likely to involve emotionally difficult material and usually involve two opposing parties in conflict, features that can reduce a litigant's ability to process information [29]. Common court forms help tenants respond to eviction actions, domestic violence survivors get restraining orders against their abusers, and parties in divorce actions resolve disputes over the custody of their children. Lack of money, and the difficulties of navigating the bureaucratic hurdles that come with a low income life in the United States, may further burden litigant processing speed and accuracy in completing forms [8].

2.2.1 How substance affects litigant accuracy in completing forms

Not all responses on a form are alike. In Forms That Work, Jarrett and Gaffney [14] propose a framework for classifying form responses as follows:

- **Slot-in** responses, which can be provided without thinking, such as name and address.
- **Gathered** responses, which require the reader to spend some time locating and then entering the information that is still readily available, such as a driver's license number.
- **Third-party** responses, which require the reader to provide information that is in another person's control. For example, the income of a household member.
- **Created answers**, which require the reader to create a new response, draft a narrative, or choose among options that they had not previously considered.

Slot-in answers are the simplest to provide in complete and accurate detail. Users may face difficulty transcribing "gathered" answers accurately (e.g., they might transpose two digits when typing an ID number), and both gathered and third-party responses may be impossible to obtain. Allen Russell "Rusty" Boehm's "Ohio Method" approach to enforcing the Ohio Forms Burden Reduction Act [2] describes a similar framework to that of Jarrett and Gaffney, although he uses the terms "Standard information," "Semi-standard information," and "Limited Access Information," and a catch-all "Other."

Litigant errors in both accuracy and completeness are most likely when providing created answers. Creating an answer requires the user to:

- Read the instructions
- Recall the information or facts that will be needed to create the answer
- Accurately apply the instructions (which may be a legal rule) to their facts.

Consider this "created" response:

[ ] Do you want a jury trial?

It requires the litigant to check or not check a single box. The format of the answer could not be simpler. But the litigant may have never considered that a trial was a possibility, let alone have an opinion ready to provide. We can imagine the litigant's thought process goes something like this:

"What is a jury trial? Oh, I guess it's like on Law and Order. But what does that mean? I don't really want a trial, I just want someone to give me a restraining order. If I say 'no' does that mean I get the decision quicker?"

The litigant may be stuck thinking about this apparently simple question for quite a long time. (In fact, on an answer to eviction guided interview created by one of this paper's authors, Google Analytics showed that this question stumped many tenants facing eviction).

This second example is drawn from a Massachusetts restraining order petition:

"AFFIDAVIT: Describe in detail the most recent incidents of abuse. The Judge requires as much information as possible, such as what happened, each person's actions, the dates, locations, any injuries, and any medical or other services sought. Also describe any history of abuse, with as much of the above detail as possible." [36]

The user is provided a full blank page to provide the requested response.

As we will discuss later, this response requires the litigant to recount traumatic details of an event that may affect memory, processing speed, and accuracy. The format of the question also requires the litigant to make many choices about the level of detail, format, and structure of their response. It would be easy for a litigant to leave out important information. In comparison, the domestic violence restraining order petition in Washington State [28] separates the narrative into 7 individual questions. These include sections asking the litigant to describe recent incidents of abuse, past incidents of abuse, medical treatment, suicidal behavior, substance abuse (with detailed checkboxes), and the effect of the abuse on minor children. The Washington petition also provides the litigant a checklist of supporting evidence.

Within the emotionally burdened context of high-stakes litigation, form authors must carefully choose the proper input type for each question to maximize ease of use, and should consider replacing some long, open-ended questions that require a lot of effort to respond to with fact-oriented questions that a litigant can respond to more automatically. The litigant should be spared from the task of synthesizing, organizing, and structuring a long narrative response when the form author can easily break a long question into smaller sections.

In addition, because even simple questions place some burden on the litigant, form authors should consider removing questions that are not required for the fact finder's decision or that can be obtained by the court from an existing data source.

2.2.2 How format and input selection affect litigant accuracy in completing forms

Form completion ease is affected by both the choice of inputs that the form author made and the layout and organization of the fields on the page.

Input selection

Common response styles on written forms include:

- Short answer text fields
- Inter-lineal text responses
- Long answer text fields
- Check boxes
- Radio buttons (exclusive checkboxes)
- "Circle one" fields

These input styles range in difficulty. Checkboxes are easy to mark, although too many choices or an incomplete list can make them challenging to answer correctly. Short free-text responses make more sense than checkboxes when the user's choices can cover a very wide range of correct answers. Longer narrative answers can be time consuming for the user to interact with, but they are appropriate when the fact finder needs unstructured responses, such as "why" and "what happened." Longer narratives also give the litigant a chance to tell their story. "Circle one" inputs are uncommon and can therefore be confusing on a paper form and should be replaced with checkboxes. Similarly, interlineal text responses, such as "The Defendant is the child's _____ (mother or father)" are uncommon and should be avoided. On a printed form, radio button can be inherently confusing, unless limited to choices that exclusively describe a party, such as "Plaintiff" and "Defendant" or "Male," "Female" and "Nonbinary."

Format, order, and layout of fields on the page

Input choices are not the only choice that a form designer can make that affects the ability of a litigant to complete the form: layout, density of fields on the page, use of whitespace, and logical grouping are also important. In addition, the use of appropriate capitalization [18] can influence reading speed and comprehension. We have little to say about font size, as the optimal font size for readability has been described with as wide a range as 9 to 18 points [11,18,21]. Some of these format choices are easy to measure, and some are difficult.

A consistent brand identity, with shared headings and a familiar layout across forms will reduce the litigant's effort to locate and provide an appropriate response to each question.

Field density, as a proxy for whitespace [20], is also important. Form authors sometimes try to fit a form onto a small number of pages, at the expense of readability. Placing too many fields on a page can reduce the litigant's ability to fit their response on the form as well as their ability to locate the most important information on the form. Correctly used, whitespace can also provide semantic grouping of information [20].

Finally, grouping questions in a logical order can affect form completion time. Jarrett and Gaffney [14] discuss the negative effect on completion rates of having questions in a surprising order. This is particularly true of questions that require the litigant to gather information from the same source. We suggest form authors read their forms carefully to make sure that like fields are grouped together.

2.3 Measuring burden on form users

The Paperwork Reduction Act of 1995 [31] in the United States tasked the United States Government with minimizing form completion **burden** on the person filling out the form. Because easy to read forms can still be overly long, burden is best considered independently from other complexity metrics. Burden is usually measured as a function of the respondent's cost in time and money [27,37]. Time burden is the most relevant for court forms, and may include the time it takes the user to:

- Read the form
- Gather information
- Respond to the form

In addition, while many forms that users interact with request routine, unemotional facts, court forms often are centered on a traumatic or high-conflict experience. Therefore, when focusing on court forms, it is also important to consider the psychological burden imposed on the user.

2.3.1 Measuring time burden

The time it takes to complete a form depends on the time it takes to:

1. Read any instructions and field prompts.
2. Write down the response once it has been retrieved or created.
3. Retrieve or create a response to each field.

Trauzettel-Klosinski et. al. [26] measured reading speed across 17 languages using standardized text, and found that across populations the average reading speed was 184 words per minute. Reading speed is much slower than the average for significant sub-populations, such as those with dyslexia [17]. Our formula assumes a reading speed of 150 words per minute to account for population variation and to reach something more than a bare majority of readers. Average handwriting speed has been measured at 40 characters per minute [30].

We assigned a "time to answer" to each classification of field in the Jarrett and Gaffney framework of slot-in, gathered, third-party and created fields [14]. Both the time to answer and the distribution are a simplified estimate based on our collective experience working with low-income and self-represented litigant populations. We assume a normal distribution of answer times. We selected the times in table 1 to approximate answer time for each answer type.

The time to create an answer is then added to the length of the field in characters divided by the average handwriting speed of 40 characters per minute, after categorizing the field length into one of the following buckets:

1. One-line answers (assuming a typical line length of 80 characters), which we assume require about 1 word of writing

2. Short-answer questions, which we round up to 2 lines
3. Medium answer questions, which we round to 5 lines
4. Long answer questions, which exceed 5 lines of text and we round to 10 lines

Table 1: Time to produce for each answer type

Answer type	Mean time to produce	Standard deviation
Slot-in	.25 minutes	.1 minutes
Gathered	3	2
Third-party	5	2
Created	5	4

We normalize answer length into these buckets rather than directly using character count because we assume that the blank space on the page is likely to be constrained by court considerations, such as limiting the number of pages that the form will use, rather than directly providing information about the actual average answer length.

To account for the variability in time to answer, we assume a normal distribution, run a Monte Carlo simulation for the time to answer each individual field and then sum the simulated values. The RateMyPDF time to answer metric is shown separately from the form's overall complexity score.

2.3.2 Measuring emotional burden

Legal forms may require litigants to disclose details of traumatic personal events. Most litigation involves conflict between two parties, and this conflict can be very personal and involve elements such as domestic violence, abusive landlord-tenant relationships, and more. Trauma has a direct link to processing speed and cognition [29]. Form authors should carefully consider the benefits and drawbacks of asking a litigant to recount such detailed information, and ensure that when it is requested it has a direct corresponding benefit to the litigant in helping obtain the relief that they requested.

Because the emotional burden of completing the form cannot always be eliminated or reduced, we do not use emotional burden as an element of our complexity score. Disclosing the traumatic event is often, but not always, central to getting relief based on that event. For example, while it is appropriate for the court to ask for an affidavit recounting an episode of abuse to grant a restraining order, it is not appropriate to ask the litigant to repeat that information in a purely administrative form that the litigant can use to enforce an order that has already been granted. These distinctions are difficult to make in an automated way.

We use the Spot [25] NLP classifier to create a first guess about the form's classification using the LIST (Legal Issues Taxonomy) taxonomy of legal problems [38]. This classification can be a useful signal as to the form's emotional burden.

3 The state form dataset

We started by surveying 50 States and the District of Columbia to identify a location on an official court website that listed standardized court forms (an initial version of this list of state court form pages was graciously shared with us by the Stanford Legal Design Lab). Some states do not have an official website that lists forms or only have forms in Microsoft Word format. Other states do not have any state-wide standardized forms or have a small number. At the low end, Louisiana had about 5 state-wide forms while at the high end, California had 1,500.

States that lack state-wide forms may have forms that vary by judicial district or county. With the exception of a small number of county-level forms in Florida, we did not include these "local" form variations in our survey.

Ultimately, we gathered PDF forms from 46 states and the District of Columbia. We have created a website that allows form authors to explore all the forms that we have collected and processed in one place, called the Form Explorer [39].

3.1 Method of gathering forms

Forms were gathered by scraping, primarily with Python's requests[40] library and custom logic. While we tried multiple approaches to scraping court websites that required varying levels of effort, we finally landed on an approach that simply crawls the main form index page and a set depth of cross-linked pages on the same domain for PDF files and downloads them all. Within FormFyxer we later apply GPT-3 to the full text of the form to create a title and a description and use the Spot NLP classifier [25] to assign it a category, information that might otherwise be obtained from more hand-tuned manual scraping. We also ask GPT-3 to write the description and summary of the form's text at a 6th grade reading level. This flexible approach to scraping court forms will allow us to keep our dataset up to date with minimal effort. The automated plain language summaries may also prove useful to the work of court staff who maintain libraries of hundreds of forms.

3.2 Results from our form benchmarking

We ran a large subset (about 15,000) of the 24,000 forms we gathered through our FormFyxer tool. We excluded some forms that could not be automatically processed, appeared to be in a language other than English, or were obtained too late in our process. This subset represents forms from 24 states. Ultimately, we obtained benchmark scores in table 2 from that dataset.

We notice that there is a lot of variability in the dataset, for almost every measure. Within a single standard deviation, for example, we range from 20 fields to less than 1 field per page. This is likely due to jurisdiction variations as to whether there are cover pages with instructions. This is an area that deserves further close attention.

The average reading grade level of almost 10th grade likely understates the difficulty of reading the form labels and prompts, due to the lower accuracy of readability scores as applied to forms, but is well above the target reading grade level of 6th

grade. There is a high observed percentage of difficult words and a high percentage of sentences written in the passive voice. Almost 13% of words in the average court form were outside of the Dale-Chall word list of the most common 3,000 English language words, while 24% of sentences were written in the passive voice. Legal citations were observed less often in the dataset, but we note that the EyeCite [4] library is not yet capable of reliably detecting state-specific short form citations.

Table 2: Form Benchmark Scores

Measure	Mean	Standard Deviation
Complexity score	23.46	10.71
Time to answer	37 minutes	87 minutes
Reading grade level [34]	9.7	3.07
Page count	2.76	4.54
Field count	48.35	80.23
Fields per page	20.03	20.15
Normalized character count per field	7.83	2.10
Sentences per page	12.35	7.81
Difficult word percent	12.86%	4.9%
Passive voice sentence percent	29.88%	22.96%
Citation count	1.22	4.37
Percent of words in all capital letters	8.29%	7.70%
Slot-in field percent	63.81%	27.14%
Gathered field percent	31.36%	24.09%
Third party field percent	0.00%	0.00%
Created field percent	0.002%	0.15%

Forms in our dataset ask a lot of litigants, with the mean page containing 20 separate fields, and the mean form asking the litigant to provide 48 separate pieces of information. We detected fewer third-party and created fields than expected, indicating we are undercounting those, but the difference between third-party, created, and gathered fields is relatively small in our complexity score and time to answer score. Our current formula estimates that the mean form requires a litigant to spend 37 minutes and the mean plus one standard deviation is 2 hours.

3.3 RateMyPDF score correlation to expert ratings

After building the initial version of the RateMyPDF score, we selected a random subset of forms and assigned them to a panel of 6 nationally recognized expert reviewers who variously work in the field of plain language, participate in form committees, and regularly work with self-represented litigants. We asked experts to rate a form's complexity on a scale from 1 to 5. Reviewers rated between 20 and 35 forms each, and each of 40 forms was reviewed by at least 3 raters.

We found statistically significant intraclass correlations among experts and between the average human rating and that of RateMyPDF. We normalized all scores for each reviewer and for RateMyPDF before further processing. The expert reviewers showed agreement with each other about which forms were complex (ICC1 0.3139, p-value=0.02), and the RateMyPDF score correlated with the average expert rating (ICC3 0.5861, p-value=0.00).

Additionally, when treated as a seventh reviewer the RateMyPDF score improved the groups agreement (ICC1 0.3931, p-value 0.00). These results were significant ($p<0.05$), suggesting that both human and RateMyPDF ratings perform better than a random number generator at assigning a complexity score to a form. We can reject the null hypothesis that expert human and machine ratings act as random number generators. Human raters assigned scores with more agreement than would be expected by chance, and RateMyPDF scores look more like these human raters than expected by chance. Details of our analysis can be found in the FormFyxer GitHub repository [33].

4 RateMyPDF

RateMyPDF, available at https://ratemypdf.com, is a Python + FastAPI website that allows a user to upload a single PDF with form fields and obtain a variety of statistics, including a "complexity score" that compares the PDF to our benchmark dataset.

The code is split into two repositories that are available on GitHub: FormFyxer [33], and RateMyPDF [32], which is the FastAPI frontend to the Python modules contained in FormFyxer. The FormFyxer library incorporates work from several existing open source projects, including EyeCite [4], PassivePy [23], scikit-learn [19] and spaCy [13], and leverages the commercial GPT-3 large language model for some additional machine learning tasks (specifically, text summarization), as well as the Suffolk Legal Innovation and Technology Lab's NLP issue spotter, Spot [25].

While RateMyPDF currently operates on a single form at a time, the underlying FormFyxer library was created for bulk processing of PDF forms. We used it to obtain our benchmark scores from the full state dataset, but it can also be used to rank and compare forms within a single jurisdiction. We display aggregate scores for jurisdictions in our companion website, the Form Explorer [39].

RateMyPDF has a simple interface. The first screen prompts the user to upload a Microsoft Word or PDF file; once the file is uploaded, the website displays both summary and detailed statistics that compare the form to our benchmark form set. In addition, a number of suggestions are displayed to help authors improve the form, including suggested word substitutions from both U.S. and U.K. official plain language sites.

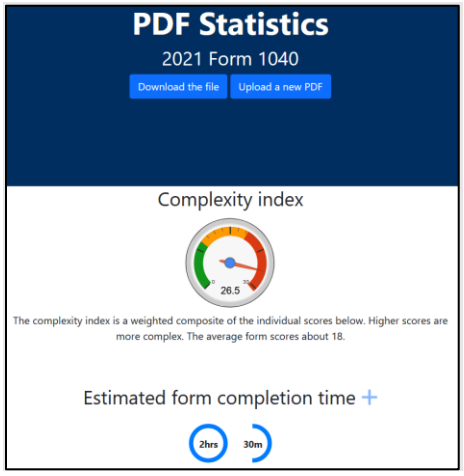

Figure 1: screen capture of RateMyPDF.com

As of the writing of this paper, RateMyPDF measures 14 features of a form. Those features are then weighted and aggregated into a single score. The score can then be compared to our benchmark set of 24,000 forms.

The features we measure are:

- Reading grade level (a consensus score)
- Percent of difficult words (currently drawn from the Dale-Chall word list)
- Use of calculations
- Number of pages
- Number of legal citations per field
- Average number of fields per page
- Normalized answer length per field
- Sentences per page
- Percent of passive voice sentences
- Percent of words written in all capital letters
- Percent of "slot-in" fields
- Percent of "gathered" fields
- Percent of "third-party" fields
- Percent of "created" fields

These features represent what we have identified as the most important non-correlated features of a form.

4.1 NLP-based field normalization and classification

Our complexity score relies on our field name normalization and our automated classification of normalized fields into "slot-in," "gathered," "third-party" and "created". We trained a proof of concept ML model using traditional classification techniques to assign normalized field names based on several features including: (1) the name of a field based on adjacent text; (2) the normalized name of the previous field; (3) the relative location of the field on the form; and (3) the topic of the field as identified by Spot[25]. This model is combined with simple heuristics to assign a meaningful label to each field in PDF forms. We normalize field names so that they are snake case (lower case words separated by a "_" character), under 30 characters in length, and where possible so that they match the Suffolk LIT Lab's Document Assembly Line standard for PDF field labels [41]. Early experiments with GPT-3 suggest adding an LLM to our heuristics may improve our automated labeling.

Fields that match the Assembly Line standard have a known semantic content and can be classified with heuristics. For example, keywords, such as "address," are used as a signal combined with the NLP model to determine that a field prompting for the litigant's address is a "slot-in" field that requires little time to prepare a response.

In addition to helping with the classification of the fields by type, these standardized labels allow interactive legal application authors to save significant time and help standardize the interactive legal applications built around these forms, facilitating processes like that our lab applied to automate dozens of forms during the Covid-19 pandemic in Massachusetts [24].

We built a pipeline around open source tools to facilitate this field normalization. As we gathered our state form dataset, we noticed that there was a lot of variability in the contents of the PDFs. Some had existing form fields and labels added. Some of those labels appeared to be added using Adobe Acrobat's "recognize form fields" function. Some were added by hand. And some were complex, nested structures that were not recognized using the most common open source PDF libraries. Some states had many forms in the Adobe LiveCycle (XFA) format. The XFA format is proprietary to Adobe and was not compatible with our field normalization tool. We built a small pipeline to convert XFA forms to standard PDFs. No existing open source tools offered this feature.

To address the remaining lack of standardized forms, we built several PDF manipulation functions into the FormFyxer [33] library. Those new functions include an auto field recognition function that works similarly to the Adobe Acrobat function. We use the well-known computer vision tool OpenCV [3] to identify boxes and lines on the PDF, as well as searching the text of the PDF for checkboxes created by using an open and closing square bracket: "[ ]". Either a checkbox or text field is added at the location of the identified "blank" space on the form. A draft field name is then created by gathering text that surrounds the field and using the field normalization model to create an automatic summarization of the full text.

4.2 Use of GPT-3 large language model

We use GPT-3 in three ways: (1) to help identify the nature of data and discard improperly formatted or irrelevant data, (2) to extract metadata from existing text, and (3) to summarize existing text.

GPT-3 performs well at identifying poorly formed source data, allowing us to save time that would otherwise be required to clean and correct large sets of PDFs with conditional prompts. For example, to obtain a plain language name for each form, we

provide GPT-3 with a prompt that contains the full text of a PDF followed by "If the above is a court form, write the form's name, otherwise respond with the word 'PoorlyFormedForm.'" This conditional prompt also limits the number of GPT-3 API calls that are required.

The third use we make of GPT-3 in our project is to transform the source data with a prompt to summarize and rewrite it at a 6th grade reading level. This output is presented to the end-user as a set of suggestions, allowing them to double-check the tool's work. This use, anchored to the source data, reduce the risk of LLM's known tendency to "hallucinate," or provide factually incorrect responses.

4.3 How recommendations are presented

When working on RateMyPDF, we relied on frequent workshopping of the tool with potential users. We shared and presented it with a group of 8 legal aid providers who meet with our team weekly to build interactive legal applications. We have workshopped early versions of the tool with Michigan court staff who are implementing a wide-scale form simplification project. We also discussed the project with staff at Pew Charitable Trusts and early versions of the guidelines were shared for review with document automation experts who participate in Law Help Interactive's monthly trainings and panels. A consistent request from our reviewers was to add easily implementable suggestions along with the statistical information.

RateMyPDF currently makes the following suggestions:

- Each component of the score is listed separately, together with a mean and a standard deviation from our larger dataset of forms.
- We use GPT-3 to provide a plain language draft of the form's name and a summary of the form based on the form's text.
- We identify citations with EyeCite [4] and suggest removing them from the document.
- We identify sentences that contain passive voice with the PassivePy library [23] and highlight the "passive" portion of the text. PassivePy also leverages the spaCy NLP tool to classify sentences as passive or not.
- We list words that do not appear on the Dale-Chall wordlist
- We highlight suggested replacements for complex terms that appear on the U.S. government's plainlanguage.gov [42] site
- We suggest replacement of gendered terms with gender neutral alternatives

One limitation that we observed with the use of the EyeCite [4] citation extractor is that it performed best on citations to federal case law and reported decisions. It did not identify state short-form citations common on legal forms, which leads to citations being undercounted.

Davison and Kantor [6] observe that a formula that measures readability has limitations for improving the readability of the texts it has measured. The features it measures can be accurate in naturally written text, but as soon as an author works to improve the score, they may reach for fixes that fool the instrument without improving the text's readability. Including specific recommendations for improvements that will improve the form's usability without "fooling" the algorithm can reduce this risk.

4.4 Using RateMyPDF to compare forms across jurisdictions

RateMyPDF allows for real-time evaluation of individual forms. We have built a companion website, called the Form Explorer [39], which uses our full 24,000 form dataset to allow court form authors to search and compare forms across jurisdictions. For example, if a court in Michigan is building a new fee waiver petition, they can use the Form Explorer website to locate semantically similar forms in other jurisdictions and compare them across several dimensions. The forms in our Form Explorer website are classified by issue type with Spot [25] and have field names that are normalized with the ML model within FormFyxer. One insight we hope form authors can obtain from this information is to identify which fields are common across jurisdictions and which ones are unique. We expect that this information can help form authors support an argument for process simplification in their jurisdictions.

We hope to combine the form comparison feature in the Form Explorer with RateMyPDF so that a form author can compare an arbitrary PDF with forms in other states.

4.5 Comparing to prior work

When researching existing examples of automated improvement of administrative forms that looked beyond traditional readability measures, we discovered AMesure, a web platform that evaluates and offers suggestions to improve French-language administrative texts [10]. Like RateMyPDF, AMesure uses a statistical approach that leverages language models when scoring text rather than the mechanical approach in readability measures like Flesch-Kincaid and Dale-Chall. However, AMesure is aimed at texts, not forms, and it does not purport to measure burden, only readability.

5 Directions for future research

5.1 Assigning a target score

Currently, RateMyPDF reports a complexity score for each form, but the score is value neutral. We provide the context of where the form is in comparison to the mean and standard deviation for the full population of state forms. From our experience working with self-represented litigants, we find it likely that the "ideal" form is less complex than the mean form. We have started this work by asking our panel of 6 experts to assign both a complexity score and a value judgment about how "good" the form is. Interestingly, answers to this question from our experts were much more varied than the responses to our question about how complex each form was, but we hope to eventually be able to assign something like a letter grade (A-F) to each form. This might allow us to

meaningfully group and compare sets of forms for ease of use without false precision.

5.2 Refining estimates with real-world benchmarking

After creating our formula that provides a time to answer for a form, we learned of research by Dr. Cyprian Ejiasa [2] that analyzed real-world timing for completing 124 representative government forms in Ohio. Those times are in table 3.

Table 3: Time to complete from real-word timing

Answer Type	Low	Medium	High
Standard information	1.02 minutes	1.74	2.46
Semi-standard	0.82	1.11	1.40
Limited access	5.29	5.65	6.01
Other		2.83	

Our estimated times in table 1 correspond roughly to the real-world figures discovered in Dr. Ejiasa's research in 1980. It would be useful to revisit this research and obtain an updated benchmark with real users on high-stakes court forms. We note that some features of court forms are relatively unique, such as long narrative responses and affidavits.

5.3 Large Language Models as a tool for directly improving readability of text

We have had promising early results when using GPT-3 to re-write sentences that are complex with simple prompts like "Write the following at a 6th grade reading level," and we make use of similar prompts when asking GPT-3 to summarize text. As instructors who have taught plain language techniques to law students and recognize the difficulty that students face with translating complex legal topics into clear writing at a 6th grade level, we feel this may be an important future task for GPT-3 that deserves further investigation. We are considering a cost effective and responsible way to integrate a redraft of the form with GPT-3 rewritten sentences into the recommendations provided by RateMyPDF.

5.4 Metrics to consider for the RateMyPDF score

We asked our panel of experts for their thoughts on the measures that we included in RateMyPDF. These give us good direction to consider re-weighting the current metrics. A future version of RateMyPDF might benefit from including:

- A direct measure of whitespace. We will consider using the OpenCV library for this task.
- A measure of field ordering. One approach we are investigating is to measure distance from a grouping created by GPT-3. Early experiments are promising.
- A different word list. The Dale-Chall list, created in 1948 with texts for children and updated in 1995 (and omitting words like divorce, tenant, and email), is not fully representative of difficult words in modern court forms read by adults. We plan to investigate the use of our dataset of the text from 24,000 court forms to create a more tailored list that can be tested with self-represented litigants.

5.5 Extension to guided interviews

Prior to beginning the work that led to RateMyPDF, our lab was focused on building guided interviews with the Docassemble web framework. We built a simple tool to analyze Docassemble interviews, but realized that a tool that analyzed printable forms would have broader use. A valuable future project would be to combine these tools.

6 Conclusion

Given the vast quantity of standardized legal forms in the United States, form simplification can be a daunting task. Washington State spent almost a decade simplifying its official legal forms. Michigan is currently in the middle of a two-year engagement with consultants to improve the readability of its legal forms, and the project will likely end with a subset of the forms that will be models for the court's form authors to continue to simplify on their own. When applied in batch to a court's library of forms, tools like RateMyPDF can help enforce standards, give clear direction to revise forms, benchmark progress, and focus the court's efforts on the forms that will provide the greatest payoff. RateMyPDF can scale the court's efforts to help self-represented litigants with simpler forms.

Yet form simplification can only go so far. Court form authors are constrained by the square corners of a piece of paper, and often try to avoid more than 2 or 3 pages for a single form. When designing a form for a legal problem with several options, such as restraining orders that depend on the relationships of the parties or a divorce proceeding that may or not involve children and accompanying custody decisions, form authors need to choose whether to include long pages of instructions and whether to combine or separate forms. Tradeoffs between comprehensive help, ease of locating the proper form, and overwhelming litigants can be complex. Interactive legal applications solve these problems. Branching logic, just-in-time instructions and context can improve the accuracy of form completion. Conditional text can make one form into many, simplifying the litigant's task in selecting the correct form.

RateMyPDF and the FormFyxer library are only a small piece of an ecosystem of tools that our lab is building. We have used the FormFyxer library to normalize fields and extract information from PDFs that can then be used to build draft automations. The normalized fields are linked to standardized questions. This ability to speed up automation may end up being the most important way that RateMyPDF can improve access to justice, but better court forms are an important first step.

ACKNOWLEDGMENTS

Michelle Bernstein, Caroline Robinson, and Lily Yang each gave generously of their time to help identify key features of forms that

affect difficulty in completion. Our expert reviewers were Laurie Garber, Marc Lauritsen, Maria Mindlin, Josh Lazar, Matthew Newsted, and Christian Noble.

We would also like to thank Caroline Jarrett and Gerry Gaffney whose text Forms That Work was an important inspiration for our framework to classify fields by answer type.

REFERENCES

[1] Rebekah George Benjamin. 2012. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educ Psychol Rev* 24, 1 (March 2012), 63–88. DOI:https://doi.org/10.1007/s10648-011-9181-8

[2] Allen Russell Boehm. Ohio Forms Burden Reduction Act. Ohio (on file with author).

[3] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[4] Jack Cushman, Matthew Dahl, and Michael Lissner. 2021. eyecite: A tool for parsing legal citations. *JOSS* 6, 66 (October 2021), 3617. DOI:https://doi.org/10.21105/joss.03617

[5] Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Readability: Instructions. *Educational Research Bulletin* 27, 2 (1948), 37–54.

[6] Alice Davison and Robert N. Kantor. 1982. On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations. *Reading Research Quarterly* 17, 2 (1982), 187–209. DOI:https://doi.org/10.2307/747483

[7] William H. DuBay. 2007. *Smart Language: Readers, Readability, and the Grading of Text*. Retrieved February 3, 2023 from https://eric.ed.gov/?id=ED506403

[8] Anne Fernald, Virginia A. Marchman, and Adriana Weisleder. 2013. SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science* 16, 2 (2013), 234–248. DOI:https://doi.org/10.1111/desc.12019

[9] Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, (1948), 221–233. DOI:https://doi.org/10.1037/h0057532

[10] Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. AMesure: A Web Platform to Assist the Clear Writing of Administrative Texts. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Suzhou, China, 1–7. Retrieved November 9, 2022 from https://aclanthology.org/2020.aacl-demo.1

[11] Dr Jörg Fuchs, Tina Heyer, and Diana Langenhan. 2008. Influence of Font Sizes on the Readability and Comprehensibility of Package Inserts. *Pharm. Ind.* (2008).

[12] Paula Hannaford, Scott Graves, and Shelley Spacek Miller. 2015. *The Landscape of Civil Litigation in State Courts*. National Center for State Courts. Retrieved May 1, 2023 from https://www.ncsc.org/__data/assets/pdf_file/0020/13376/civiljusticereport-2015.pdf

[13] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Retrieved February 2, 2023 from https://spacy.io/

[14] Caroline Jarrett, Gerry Gaffney, and Steve Krug. 2008. *Forms that Work: Designing Web Forms for Usability* (1st edition ed.). Morgan Kaufmann, Amsterdam ; Boston.

[15] Marc Lauritsen and Quinten Steenhuis. 2019. Substantive Legal Software Quality: A Gathering Storm? In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ACM, Montreal QC Canada, 52–62. DOI:https://doi.org/10.1145/3322640.3326706

[16] Irving Lorge and Raphael Blau. 1941. Reading Comprehension of Adults. *Teachers College Record* 43, 3 (December 1941), 1–6. DOI:https://doi.org/10.1177/016146814104300303

[17] Shelley Miller-Shaul. 2005. The characteristics of young and adult dyslexics readers on reading and reading related cognitive tasks as compared to normal readers. *Dyslexia* 11, 2 (2005), 132–151. DOI:https://doi.org/10.1002/dys.290

[18] A. Miniukovich, A. De angeli, S. Sulpizio, and P. Venuti. 2017. Design guidelines for web readability. In *DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems*, Association for Computing Machinery, Inc, Edinburgh, 285–296. DOI:https://doi.org/10.1145/3064663.3064711

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, (2011), 2825–2830.

[20] Janice Redish. 2000. Readability formulas have even more limitations than Klare discusses. *ACM J. Comput. Doc.* 24, 3 (August 2000), 132–137. DOI:https://doi.org/10.1145/344599.344637

[21] Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make It Big! The Effect of Font Size and Line Spacing on Online Readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), Association for Computing Machinery, New York, NY, USA, 3637–3648. DOI:https://doi.org/10.1145/2858036.2858204

[22] John Sabatini. 2015. *Understanding the Basic Reading Skills of U.S. Adults: Reading Components in the PIAAC Literacy Survey*. ETS Center for Research on Human Capital and Education. Retrieved February 3, 2023 from https://eric.ed.gov/?id=ED593006

[23] Amir Sepehri, David Matthew Markowitz, and Mitra Mir. 2022. PassivePy: A Tool to Automatically Identify Passive Voice in Big Text Dat. DOI:https://doi.org/10.31234/osf.io/bwp3t

[24] Quinten Steenhuis and David Colarusso. 2021. Digital Curb Cuts: Towards an Open Forms Ecosystem. *Akron Law Review* 54, 4 (2021), 2.

[25] Suffolk Law School's Legal Innovation and Technology Lab. About Spot. Retrieved February 9, 2021 from https://spot.suffolklitlab.org/

[26] Susanne Trauzettel-Klosinski, Klaus Dietz, and the IReST Study Group. 2012. Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST. *Investigative Ophthalmology & Visual Science* 53, 9 (August 2012), 5452–5461. DOI:https://doi.org/10.1167/iovs.11-8284

[27] Linda Veiga, Tomasz Janowski, and Luís Soares Barbosa. 2016. Digital Government and Administrative Burden Reduction. In *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance* (ICEGOV '15-16), Association for Computing Machinery, New York, NY, USA, 323–326. DOI:https://doi.org/10.1145/2910019.2910107

[28] Washington Law Help. 2022. How to File Petition for Order of Protection. Retrieved February 6, 2023 from https://www.washingtonlawhelp.org/files/C9D2EA3F-0350-D9AF-ACAE-BF37E9BC9FFA/attachments/9100D6C9-D107-4B15-87B3-A898F12B6FD8/3701en_how-to-file-petition-for-order-of-protection.pdf

[29] Antoinette Welsh. 2013. Effects of Trauma Induced Stress on Attention, Executive Functioning, Processing Speed, and Resilience in Urban Children. *Seton Hall University Dissertations and Theses (ETDs)* (December 2013). Retrieved from https://scholarship.shu.edu/dissertations/1907

[30] Jenny Ziviani and John Elkins. 1984. An Evaluation of Handwriting Performance. *Educational Review* 36, 3 (November 1984), 249–261. DOI:https://doi.org/10.1080/0013191840360304

[31] 2015. Paperwork Reduction Act (44 U.S.C. 3501 et seq.). *Digital.gov*. Retrieved February 2, 2023 from https://digital.gov/resources/paperwork-reduction-act-44-u-s-c-3501-et-seq/

[32] 2023. RateMyPDF. Retrieved February 3, 2023 from https://github.com/SuffolkLITLab/RateMyPDF

[33] 2023. FormFyxer. Retrieved February 3, 2023 from https://github.com/SuffolkLITLab/FormFyxer

[34] 2023. Textstat. Retrieved February 7, 2023 from https://github.com/textstat/textstat

[35] How to write good questions for forms - NHS digital service manual. *nhs.uk*. Retrieved February 6, 2023 from https://service-manual.nhs.uk

[36] Restraining order/abuse prevention order court forms | Mass.gov. Retrieved February 6, 2023 from https://www.mass.gov/lists/restraining-orderabuse-prevention-order-court-forms

[37] How to estimate burden | A Guide to the Paperwork Reduction Act. Retrieved November 9, 2022 from https://pra.digital.gov/burden/estimation/

[38] LIST:Legal Issues Taxonomy. *LIST: Legal Issues Taxonomy*. Retrieved February 7, 2023 from https://taxonomy.legal/

[39] About the Form Explorer? Retrieved February 7, 2023 from https://suffolklitlab.org/form-explorer/

[40] Requests: HTTP for Humans™ — Requests 2.28.2 documentation. Retrieved February 3, 2023 from https://requests.readthedocs.io/en/latest/

[41] Field labels to use in template files | The Document Assembly Line Project. Retrieved February 3, 2023 from https://suffolklitlab.org/docassemble-AssemblyLine-documentation/docs/label_variables

[42] plainlanguage.gov | Choose your words carefully. Retrieved April 29, 2023 from https://www.plainlanguage.gov/guidelines/words/