Towards Unlocking the Narrative of the United States Income Tax Forms

Esme Manandise
Intuit Futures
Mountain View, California, USA
esme_manandise@intuit.com

Abstract

The present study contributes to the literature on the language of the tax-and-regulations domain in the context of highly-formatted tax forms published by a federal agency. Content and form analyses rely on a methodology that looks for meaning and patterns in connection to the main purpose of income tax filing, i.e. figuring out calculations to determine whether taxes were overpaid or owed to the United States Internal Revenue Service. Profiling the income-tax forms by spelling out language regularities across the set has at least two advantages. Firstly, profiling contributes to the understanding of how the *2010 Plain Writing Act* mandate of '*clear and simple'* writing is being achieved—if at all. Secondly, profiling a small, unannotated corpus can help determine the Natural Language Processing approach best fitted to extract, represent, and execute automatically tax calculations expressed as arithmetic word problems.

1 Introduction

The term "narratives" refers to accounts of ideas or connected '*events*', whether factual or not, through oral or written communication. Narrative understanding and qualitative content analysis are related tasks as they study the practices, beliefs, needs, and values of groups of individuals. Other than eliciting universal lamentation—independently of one's moral view on the necessity of taxation for a civil society, the tax-and-regulations domain on the communication dimension is not popular with practitioners of discourse analysis, narrative exploration, or natural language automation. Narratives are stories and, to most, there isn't much storytelling in the tax-and-regulations domain—though a 1040 tax-return form the size of a postcard made a good *yarn.*

In the most literal sense, tax forms consist of embedded *stories* with words, phrases, sentences, fragments and tables through which run threads to output dollar amounts—as input to tax-form lines or as the final amount (refundable to taxpayer or owed to the Internal Revenue Service (IRS)). Tax forms, and their associated schedules and worksheets, provide instructions and clarifications as well as prompt taxpayers for qualitative and quantitative personal information. In addition, distributed throughout a form and across forms, are arithmetic word problems of varying complexity. To solve them, filers must understand content and handle amounts as input

Forms	Segments
F4868	Late filing penalty is usually charged if your return is filed after the due date. The penalty is usually 5% of the amount due for each month or part of a month your return is late. The maximum penalty is 25%. If your return is more than 60 days late, the minimum penalty is $210 (adjusted for inflation) or the balance of the tax due on your return, whichever is smaller
F8829	Line C times line D divided by 12 times $5.00 times line E
F1041	If line 25 is larger than the total of lines 23 and 26, enter amount overpaid.
F2441	Add the amounts on lines 12 and 13 and subtract from that total the amount on line 14.
F8941WKS	If the result is not a multiple of $1,000, round the result down to the next lowest multiple of $1,000
F8949	Add the amounts in columns (d), (e), (g), and (h) (subtract negative amounts)

Table 1: Calculations as Raw Text

Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019), pages 33–41
Turku, Finland, 30 September,

to basic operations (addition, subtraction, multiplication, division, percentage conversion, rounding). Sometimes, to complete the calculation, they must make the arithmetic operation explicit. With stacked operations, they must apply the operations in their correct order. Consider the examples in Table 1 above.

Tax forms are published by the IRS which, as a federal government agency, complies with the *Plain Writing Act* of 2010. The language in tax forms is supposedly '*clear and simple*' to help with content understanding. Simplicity should encourage filers to comply[1] with the Tax Law.

For natural language processing (NLP) tasks which consume raw text as input, the mandate '*clear and simple*' is an ideal convenience. Can we discover how '*clear and simple*' is instantiated in tax forms? Has '*clear and simple*' turned the language of tax forms, schedules, and worksheets into an unequivocally-specific language register? Does '*clear and simple*' remove semantic and syntactic ambiguities? One of our goals in referencing the notion of '*clear and simple*' is to gain exploratory insight into the language of tax forms with the ultimate purpose of using raw text as input to the automatic (no human-in-the-loop.) detection and execution of calculations by an NLP system.

In this paper, we describe a preprocessing implementation for detecting and labeling executable calculations in raw text. More specifically, we concentrate on feature-based classifications to build a profile of the United States income-tax forms set as a whole rather than per-document profiles. Ultimately, our investigation may help to assess whether '*clear and simple*' is measurable or merely a matter of opinion[2].

2 Related Work

To the best of our knowledge, there are no publications in English that detail the language and discourse of the income-tax forms in the tax-and-regulations domain. However, glossaries of tax terms are aplenty; they are made available online and/or are published by government agencies[3], private outfits[4] and international organizations[5]; some glossaries are integrated in tax and accounting software[6]. While tax terms are important as they correspond to concepts and entities in the domain, tax-and-regulations texts do not consist merely of a collection of terms. The reductionist view that to learn the tax language is to learn its terms considers the tax language a *Toki Pona*—a pidgin of sort. Terms need to be connected by relations for tax text to be coherent.

Recently, some Tax Law scholars have shown an interest in the language of taxes as it appears in IRS publications. They have focused on the federal government agency mandate to output text in a '*clear and simple*' language. Most noticeably, Blank et al. (2017) discuss instances where the IRS transformed *'complex, often ambiguous tax law into seemingly simple statements*'. Achieving language simplicity can cause a loss of information and make content less accurate. The authors summarize their findings in three categories: (1) '*contested tax law presented through language simplification as clear tax rules*', (2) '*failure to explain the tax law with possible exceptions*', (3) tax law rewording by IRS. They discuss concrete language examples. How the change from the adverb '*materially*' in Treasury regulations to the adverb '*significantly*' in IRS publications can create uncertainty in filers when determining exclusions from taxable income of gain from the sale of a principal residence.

3 Income Tax Form Set

To build the profile, we use 234 IRS tax forms, schedules, and worksheets (individual and fiduciary) for the 2017 tax year. These are published in English in PDF format (see sample in Figure 1.) The forms have a visually-complex structure consisting of a mix of raw text as free-standing paragraphs and of tables with rows and columns, headers, instructions, cautionary notes, line labels, checkboxes, input fields, etc. (see Figure 1 below.)

We use a machine-learning-based algorithm to extract raw text from the PDF-formatted files. The context of raw text (occurrence in original layout) is recorded because the text '*position*', in

[1] According to the IRS (Blank et al. (2017)), 56% of filers use third-party advisors, 34% rely on tax preparation software, and 10% of individuals file without assistance.

[2] Tax forms have a readership of around 140,000 million filers with no uniformity in educational background or English-language literacy.

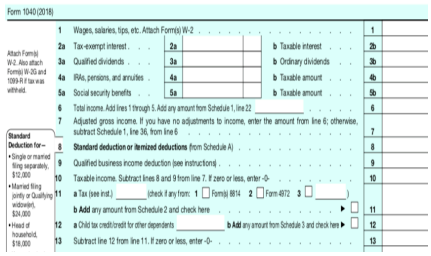
[3] For instance, IRS.gov, efile.com, psu.instructure.com

[4] For instance, law and accounting practitioners (Taxman.com or taxWorld.com)

[5] For instance, Organisation for Economic Co-operation and Development

[6] For instance, TurboTax

particular when occurring in tables, can be relevant to its interpretation.


Form 1040 (2018)

	1	Wages, salaries, tips, etc. Attach Form(s) W-2			1
Attach Form(s) W-2. Also attach Form(s) W-2G and 1099-R if tax was withheld.	2a	Tax-exempt interest	2a	b Taxable interest	2b
	3a	Qualified dividends	3a	b Ordinary dividends	3b
	4a	IRAs, pensions, and annuities	4a	b Taxable amount	4b
	5a	Social security benefits	5a	b Taxable amount	5b
	6	Total income. Add lines 1 through 5. Add any amount from Schedule 1, line 22			6
	7	Adjusted gross income. If you have no adjustments to income, enter the amount from line 6; otherwise, subtract Schedule 1, line 36, from line 6			7
Standard Deduction for— • Single or married filing separately, $12,000 • Married filing jointly or Qualifying widow(er), $24,000 • Head of household, $18,000	8	**Standard deduction or itemized deductions** (from Schedule A)			8
	9	Qualified business income deduction (see instructions)			9
	10	Taxable income. Subtract lines 8 and 9 from line 7. If zero or less, enter -0-			10
	11	a Tax (see inst.) (check if any from: 1 Form(s) 8814 2 Form 4972 3)			
		b **Add** any amount from Schedule 2 and check here			11
	12	a Child tax credit/credit for other dependents b **Add** any amount from Schedule 3 and check here			12
	13	Subtract line 12 from line 11. If zero or less, enter -0-			13

Figure 1: Income Tax Form Sample

4 A Brief Overview of the Nature of the Income Tax Narrative

The underlying schema of an income tax form narrative[7] is that of a camera-eye narration with purely matter-of-fact representation of facts, events, and actions to be taken. The text reads like a transcription with fragmentary content sequentially displayed and/or distributed across columns. The timeline between facts, events, and actions is punctuated with form-name and line references, with spatial and situational pointers like '*above*' or '*this*' as well as with temporal references such as '*current*' or '*past-due*'. Even though the narrative protagonist is referred to in the second person, the pronoun '*you*' means '*anyone*' who is filing an income tax return.

Deixis is present throughout the text of income tax forms. Deixis curates the filer's path to help complete income tax return filing. However, it is up to the filer to assign denotational meaning to deictic expressions.

And then do the maths!

5 Form Set Description and Classifications

To address the problem of the tax-form language and its embedded stories, we use descriptive statistics and classifiers with features that have immediate practical significance for the tax domain.

PDF extraction outputs structured json files wherein named fields hold various types of source data. The field of most immediate interest for our purpose is the field[8] named '*paragraph*'. Before we classify the content, we automatically segment[9] the paragraphs into a collection of individual segments. We do not use the notion of sentence, which implies the marker '*tense*' (however instantiated in the tax language). Given that content relevant to calculations may be a table header or a text fragment that points to an amount referenced by a line number, we use '*segment*' to refer to the minimal string unit used in the analysis (and as input to the NLP annotation preprocessor.) Currently, to create the tax-form profile, the NLP preprocessor only inspects content and collects information on individual segments.

5.1 A Lexical Paradigm for Feature-based Classifications

Our NLP annotation uses 2 lexical resources: (1) base lexicon for single tokens and (2) term lexicon for multiword expressions (MWE) corresponding to tax concepts and entities. The base lexicon is a repository of granular knowledge about single tokens in the domain. Many words in the base lexicon correspond to the head of a term at the phrase level, i.e. heads of terms are subject to morphological changes such as singular/plural. For instance, the single-token concept '*expense*' is the head of the MWE '*daycare expenses*' or '*research and experimental expenses*'.

Both resources have been populated automatically by mining IRS income-tax forms, schedules, worksheets, publications, and TurboTax interviews. After completion, the base lexicon was vetted by specialists.

add	{ pos:verb, arg1:obj, arg2:prep_to, arg3:prep_on, arg4:prep_through prep_thru, arg5:prep_for, semtype:arithmetic_operation, accumulation tr:arg1toarg2, syn:combine, sum, total, freqs: ……}
total	{ pos:adj, pos:noun, pos:verb, arg1:prep_from, arg2:prep_on, arg3:prep_for, semtype: arithmetic_operation accumulation amount outcome property, tgtwd:sum, syn:add, freqs: ……}

Table 2: Lexical Key-Values Pair Sample

Lexical entries in the lexica have been designed as a pair *{key:values}*. The values themselves can be of the type *{key:values}*. The '*values*' fields have been augmented with Wordnet and in-house-Wordnet-like features to describe granular

[7] In its instructions, the IRS uses the notion of narrative to describe the process for filing specific forms like F990 or F13424-M.

[8] Issues with PDF extraction are reflected in paragraph fields as text can be inaccurately split or glommed together.

[9] Segmentation relies on linefeed tags or predefined diacritics such as semi-colon or period.

morphological, semantic, syntactic, and domain-idiosyncratic properties of the keys. Consider the entries '*add*' and '*total*' (listed in Table 2 above in abbreviated format.)

Segments are tokenized; then each token is lemmatized to enable base form matching in the lexica. When matching is successful, the lexical information (*values* field) associated with the keys is retrieved. The built-in classifiers rely on these lexically-specified value features to automatically compute segment classifications as well as flag features that can be problematic to parsing such as scope of coordination or attachment points for prepositional phrases[10]. The preprocessor collates together a shallow description for each segment.

This classification strategy was adopted to generate reports, search and group segments on clusters of shared features (in abbreviated format here):

1	Segment	if more than one form 8611 is filed, add the line 14 amounts from all forms and enter the total on the appropriate line of your return.
	Features	Arith operation_0-MWE_Tensed_Coordination_Conditional_Posambiguity_PP
2	Segment	enter your 2017 regular income tax liability minus allowable credits (see instructions)
	Features	Arith operation_2-MWE-3w-2w-_Tensed_Verb-like_Parens

Table 3: Segment Feature Labeling Sample

For instance, the feature-aggregate label informs that segment 1 is an arithmetic operation with no MWE as operands and that some conditions need to be met for the operation to apply. As for segment 2, the label classifies it as an arithmetic operation with a verb-like operator *minus*. There are 2 multiword expressions '*income tax liability*' and '*allowable credit*', of 3- and 2-words, respectively; these MWE are operand candidates. In addition, there is parenthetical material that will need checking during parsing.

This labeling schema allows us to readily search the form set as a collection of segments. For instance, there are 3,970 segments labeled '*arithmetic operation*', but only 5.18% of these use '*minus*', '*plus*', or '*times*' to express subtraction, addition, or multiplication.

5.2 General Descriptive Statistics

The United States income-tax-form set for the 2017 tax year is a small collection of 234 forms. After the PDF extraction of the structured content of tax forms, paragraphs are retrieved and each paragraph is, in turn, broken down into separate segments. General details of the set are given in Tables 4 and 5 above.

Total Number of (No.)	
Individual forms	234
Paragraphs	15,294
Single segments	41,660
Single words[1]	349,146
Segments with terms	18,164
Unique words	6,424
Unique alphabetic words	4,812
Unique non-alphabetic words	1,612
Average No. single words per sentence	8.46
Average sentence length	15.82
No. terms	23,840

Table 4: General Statistics

Word	Rank	Percent
line	1	04.90%
the	2	04.04%
of	3	02.39%
and	4	02.03%
or	5	02.01%
for	6	01.81%
form	7	01.61%
from	8	01.58%
to	9	01.56%
enter	10	01.55%
if	11	01.43%
a	12	01.35%
on	13	01.28%
tax	14	01.09%
amount	15	01.06%
year	16	01.00%
you	17	01.00%
in	18	00.95%
income	19	00.83%
total	20	00.79%
your	21	00.74%

Table 5: Top 21 Most Frequent Words

The determiner '*the*', reputed to be the most frequent word in English (*OxfordDictionaries.com*), ranks only second in

[10] In this paper, we restrict ourselves to a general description of the methodology.

our form set. '*Line*' ranks first. Far from being a stopword, '*line*' is the basic structural and functional unit not only as a marker in the PDF layout of tax content, but as content-reference pointer and content holder.

The 21-top-ranked tokens offer a glimpse at tax-form activities. One can readily create a narrative—something along the lines wherein the text is about '*income*' and '*tax*' '*on/in*' '*forms*' and '*lines*' for some '*year'*. It concerns the reader '*you/your*' who is prompted to take action by '*enter*'ing '*amount*' and '*total*' ('*and, or*') when conditions are met ('*if, and, or*'). There is traffic of content '*from*' and '*to*'.

5.3 Terms as Text Instances of Tax Concepts and Entities

Multiword expressions (MWE) or terms are terminological units which denote concepts and entities in a domain. In the tax-and-regulations domain, terms can be compositional (Nunberg et al., 1994, Baldwin, 2006) in meaning and/or in form like '*timely estimated tax payment*'; others are not like '*married filing jointly*'; yet others are mixed instances of compositionality such as '*taxable sick leave pay*' or '*cannabis duty payable*'.

The domain-term lexicon is the result of the prior task of identifying, given the domain corpus, the domain-relevant concepts and entities by means of co-occurrence/collocation-based surface statistical measures. In addition, linguistic filters delete ill-formed term candidates from the final term list. We retain only nominal terms. Table 6 provides a breakdown for the number of MWE occurrences per segment.

Filing taxes requires understanding the concepts and entities being considered, i.e. what these MWE/terms denote in the tax-and-regulations domain. About 43.6% of all segments include at least one term. And about 35% of these segments consist exactly of just terms. For instance, the segments '*net operating loss deduction*' or '*tentative income distribution deduction*' are the terms themselves.

Total No. of Segments		
41,660		
Total No. of Segments with		
0 MWE	23,496	57%
1 MWE	13,901	33%
2 MWE	2,901	7%
3 MWE	861	2%
4 MWE	287	.6%
> 4 MWE	194	.4%
Total No. of Segments that are MWE		
6,315		

Table 6: MWE Distribution

The raw text in tax forms is fragmentary with a prevalence of nominal expressions; the fragmentation and its instantiation with nominal phrases mirror not only its function in the visual layout of the source documents but also the piecemeal cumulative reading and building of calculations.

5.4 Segment Type

Segment-type classification exploits the semantically-based features associated with the keys in the lexica. As the ultimate goal of our tax NLP system is to interpret and execute calculations expressed in the input as raw text, the preprocessor labels each segment according to the schema in Figure 2 below (each segment must be flagged with one of the bottom labels).

Segments that are labeled '*arithmetic operation*' have explicit verbs, nouns or adverbs that identify the operations; they also express complete operations like '*Subtract lines 13a plus 13b from line 12*'. '*Non-arithmetic operation*' segments either include quantity-oriented concepts like '*business expense*' as in '*Total unreimbursed employee business expenses*', or include references to quantities as in '*Total net gain from Schedule D (Form 1041), line 19, column (1)*'. We further divide '*amount*' segments into (i) term-based '*amount*' segments like '*Total unreimbursed employee business expenses*' and (ii) '*arithmetic operand'* segments that reference content in form-units to be used as input to a calculation as in '*Enter the amount from column (c) on line 1*'. Finally, '*non-arithmetic non-amount*' are classified as either '*particulars*', '*date*', '*description*', or '*declaration*' like for '*Social security number*' or '*I hereby…*'. Only segments labeled '*arithmetic operation*' and '*non-arithmetic operation amount*' are of interest in the context of the automatic extraction and execution of raw text calculations by an NLP system.

Details of the semantic-based segment-type breakdown are presented in Table 7. Over a fourth of all segments (28%), are clearly identifiable as '*arithmetic operations*' and '*arithmetic operand*'. In addition, more than half of the segments (57.5%) are about '*amount*' as concepts instantiated by tax terms. Currently, we do not discriminate among '*amount amount*' segments.

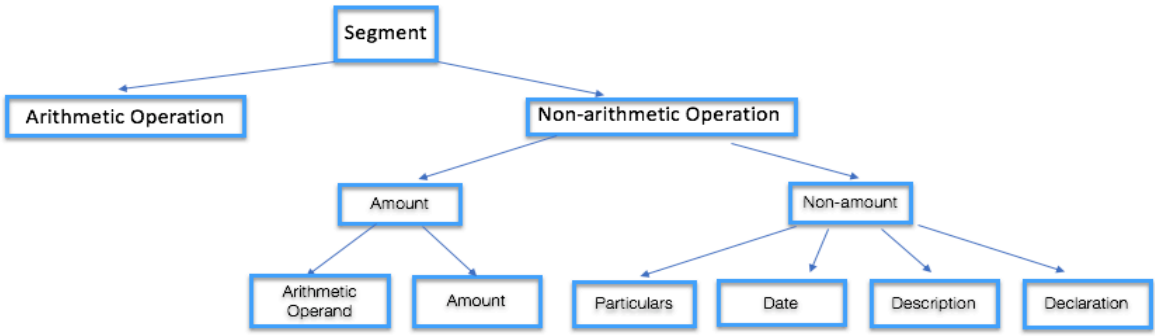


Figure 2: Semantic Segment-type Classification

Segment Total	Arithmetic Operations	Non-arithmetic Operations		
		Amount		Non-amount
		Arithmetic Operand	Amount	
	3,970	7,687		
	11,657		23,940	6,063
41,660	28%		57.5%	14.5%

Table 7: Semantic Schema for Text Calculations

A subset of '*amount*' segments function as operands to in-progress calculations like '*enter total business expenses*'. Ideally, '*enter total business expenses*' should have a label indicative of its function—*amount operand*, to distinguish it from instances where the term '*total business expenses*' is, for example, part of an explanation.

5.5 Tense Marker

Due to the nature of the highly-formatted original PDF forms, many segments are verbless. The predominant table-like layout of tax forms encourages text fragments and isolated phrases. These segments identify, label, or prefix lines and line content. We use the absence/presence of a '*tense*' marker on the candidate head of a phrase to label segments.

Excluding non-amount segments (6,063) from the total number of segments (41,660), we have 35,597 segments relevant to calculations. Of these, 58% have a noun as the head of the topmost phrase and 42% have a verb as the root node. More than half of the tense-based segments are in the imperative mode as in '*Enter here and on Form 1041, line 25b*', '*Itemize by charitable purpose*', or '*If zero or less, enter 0 here, skip lines 13 through 21, and enter 0 on line 22*'. These commands are instructions that spell out the steps to take or not to build the calculations.

The breakdown for tensed versus non-tensed segments is in Table 8.

Non-tensed			Tensed		
Arithmetic Operation	Arithmetic Operand Amount	Amount	Arithmetic Operation	Arithmetic Operand Amount	Amount
470	5,042	15,228	3,510	3,017	8,330
20,740			14,857		

Table 8: Tense-Marker Frequency per Segment-Type

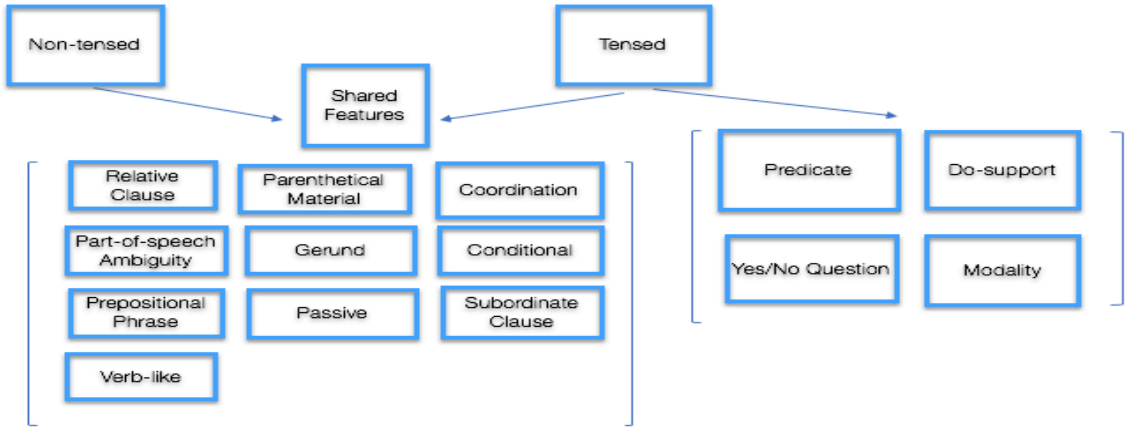


Figure 3: Linguistic Shared Features

5.6 The Structural Flavor of Segments

Automated phrase- and word-frequency lists based on the denotative classifications of parts of speech (POS) and lexical information in our lexica point to structural choices, which are used to label segments into smaller functional groups. The set of structural and functional labels that correspond to structural-syntactic instantiations in the collection of segments is shown in Figure 3. Before any deeper parsing, the labels serve as a precursor indication of the segment overall configuration.

For example, parenthetical material occurs in segments regardless of whether the segments are non-tensed as in '*ordinary income (loss) for schedule E*', or tensed as in '*if you were a real estate professional, enter the net income or (loss) you reported*.' Arithmetic operations (in full or in part) can be contained inside parenthetical material like '*enter the result as a decimal (rounded to at least three places)*' or '*Add amounts in column (i), line 26*'. The intent of the parenthetical material needs to be weighed as it may or may not be relevant to calculations as in '*Other taxable disability income (see Help)*'. 16% of all segments have text in parentheses.

6 Can You Read Me Now?

A way of determining whether the language of income-tax forms is clear and simple is to measure the text set against readability indexes. Readability measures rely on various standardized writing components like sentence length (the shorter the better), word length (the shorter the better), concrete everyday language, active voice, no legalese or jargon, tabular presentation of complex information, etc.

Conveniently, Microsoft Word (version 16.27) comes with a readability tool. Running our text set as a collection of segments abstracted away from their position in the original highly-structured PDF yields a Flesch Reading Ease of 53.5 or a Flesch-Kincaid Grade Level of 8.8. With a grade of 8-9, the text set should be understood by 13- to 15-year-old individuals. In isolation, the language of tax-form segments appears on average clear and simple. Substantially, the language complies with *Plain Writing* principles.

However, readability measures and *Plain Writing* principles largely ignore the questions of how filers make sense of fragmented texts, of how each line relates to other lines in a cumulative reading process. More importantly, while they focus on lexes and the structures of phrases, these measures and *Plain Writing* principles disregard meaning and interpretation, i.e. to understand the text the filer must determine what the terms denote in the domain. Not a trivial reading task as 43% of all segments include at least one multiword term. For instance, the structure of and each of the words in the segment '*tax on lump-sum distributions'* are common, everyday language but what does '*lump-sum distributions*' denotes in the tax world?

One of the recommendations for writing clearly and simply is to use tabular presentation of complex information. Such display result in compressing content into fragments that can be displayed in table rows and headers. Such compression results in noun compounds and nominal phrases of varying complexity (58% in our set), which can make content less explicit. *Nominalizations* may be efficient for readers who

are expert on or familiar with the tax-and-regulations domain, i.e. readers who can infer non-overt relations among concepts (and their token instantiations). These readers may be able to keep up (from line to line, from tax form to tax form) with the story on how to figure out their taxable income.

Finally, while the language of arithmetic operations expressed as text reads on average clearly and easily, understanding what the calculations consist in (operations and operands) can be challenging. Consider the following sequential segments; '*If line 27 is $186,300 or less ($93,150 or less if married filing separately for 2016), multiply line 27 by 26% (0.26). Otherwise, multiply line 27 by 28% (0.28) and subtract $3,726 ($1,863 if married filing separately for 2016) from the result.*' The Flesch-Kincaid Grade Level puts the above paragraph at 19.5, which is a level for skilled readers.

The complexity of some arithmetic operations along with concept and entity denotation in the tax domain may explain why only 10% of taxpayers file without any type of assistance.

7 Conclusion and Additional Questions

In this paper, we offer a first attempt at describing the language of income-tax forms. We viewed the task through a language analysis lens with no attempt at more logic-oriented semantic modeling. This approach also helps with the discovery of content and form that are idiosyncratic to the domain.

We discuss some basic syntactic and semantic patterns discovered through various statistical regularities across the segment set. The tabular presentation of content in the original PDF files has the effect of compressing the language resulting in a high number of noun compounds or multiword expressions wherein the relationships among concepts remain implicit as it is the case, for instance, with missing prepositions ('*living expenses*' with implicit preposition '*of*' versus '*distribution expenses*' with implicit preposition '*from*'.) Noun compounding can also introduce adjectival scope ambiguity. For example, is '*tentative*' a modifier of '*income*' or '*deduction*' in the expression '*tentative income distribution deduction*' as '*tentative income*' itself appears enough times across the tax-form set to be considered a MWE or tax concept?

Various labeling schemata, that incorporate our observations from descriptive statistics about income-tax forms, have been implemented to annotate raw segments automatically. This type of automatic annotation makes it easy to poke around segment sets (or any tax-form set, from tax year to tax year.)

A language-oriented description of how calculations or arithmetic word problems are displayed in the source PDF documents and expressed in raw text can help decide on the NLP approach and the level of analytic granularity best fitted to extract and to represent calculations so as to have these automatically interpreted and executed by downstream NLP components. For instance, is tense a linguistic feature necessary for the interpretation of calculations? What about modals? Should the structure of noun compounds and nominal phrases, a subset of which are instantiated by domain multiterm expressions, be made transparent? Is this level of granularity necessary for an automatic processing of calculations? While it may matter to human readers to have relations among members of complex nominal expressions explicitly stated to help understand tax-term-based calculations[11], it may be sufficient, i.e. '*clear and simple*', for an NLP implementation to output accurate calculations by treating these expressions as opaque nominal singletons with no internal structure.

Acknowledgments

We thank Saikat Mukherjee and two anonymous reviewers for helpful comments.

References

Joo Jung An and Ned Wilson. 2016. *Tax Knowledge Adventure: Ontologies that Analyze Corporate Tax Transactions*, in *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, Shanghai, China. June 08 2016. ACM, pages 303-311.

Timothy, Baldwin. 2006 . *Compositionality and multiword Expressions: Six of One, Half a Dozen of the Other*? In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Sydney, Australia. Association for Computational Linguistics, page 1. https://www.aclweb.org/anthology/W06-1201

[11] A topic worthy of some psycholinguistic experiments.

Joshua Blank and Leigh Osofsky. 2017. *Simplexity: Plain Language and the Tax Law*. In *66 Emory Law Journal 189*, pages 1-77. NYU Law and Economics Research Paper No. 16-17; University of Miami Legal Studies Research Paper No. 16-20

Stephen Cohen. 2005. *Words, Words, Words!!! Teaching the Language of Tax*. In *Georgetown Law Faculty Publications and Other Works*. 579. Pages 600-605. https://scholarship.law.georgetown.edu/facpub/579

Michael Curlotti and Eric McCreath. 2011. *A Corpus of Australian Contract Language: Description, Profiling and Analysis*. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*. ACM, 2011. http://dx.doi.org/10.2139/ssrn.2304652

Isabella Distinto, Nicola Guarina and Claudio Masolo. 2013. *A well-founded ontological framework for modeling personal income tax.* In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*. Rome, Italy. pages 33-42.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 636–644, Los Angeles, California, Association for Computational Linguistics

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. *Idioms*. In Stephen Everson, editor, *Language,* pages 491–538. Cambridge University Press.

Plain Writing Act Compliance Report, Internal Revenue Service. 2016. https://www.irs.gov/pub/irs-pdf/p5206.pdf

Rachel Stabler. 2014. *What We've Got Here is Failure to Communicate: The Plain Writing Act of 2010*. In *Journal of Legislation*, Vol. 40, No. 2, pp. 280-323, 2014. https://ssrn.com/abstract=2574207

The Oxford English Corpus: Facts about the language. OxfordDictionaries.com. Oxford University Press. Archived from the original on December 26, 2011. Retrieved June 22, 2011.