How to apply for financial aid: Exploring perplexity and jargon in texts for non-expert audiences

Laura Manor[1] **Yiran Su**[2] **Zachary W. Taylor**[3] **Junyi Jessy Li**[1]
[1] Department of Linguistics
[2] Department of Electrical and Computer Engineering
The University of Texas at Austin
[3] Trellis Company
manor@utexas.edu, yiransucdr@gmail.com,
zach.taylor@trelliscompany.org, jessy@austin.utexas.edu

This extended abstract presents a preliminary study of a new dataset from a genre of research that has little prior work in NLP: United States Title IV federal student aid application instructions (in English) gathered from official websites of post-secondary institutions. Financial aid communications for such colleges and universities have been highly criticized for their complexity (Dynarski and Scott-Clayton, 2008; Rosinger, 2019; Feeney and Heroff, 2013). In particular, the context-specific jargon[1] is found intimidating for applicants (Ardoin, 2013); this consequently exacerbates issues in post-secondary education accessibility between L1 and L2 speakers of English in the US (Taylor, 2020).

The identification and interpretation of genre-specific terminology have been recognized as a crucial step in the understanding of technical texts aimed at expert audiences, including in domains such as biomedical (Lee et al., 2020), engineering (Jin et al., 2018), legal (Moreno-Schneider et al., 2020), and finance (Maarouf et al., 2020). However, there is noticeably less research done on the use of such terminology in texts meant for a broad audience. This data sets includes texts which are a prime example of a genre which needs both precision and a low barrier of entry. The language used in the instructions must be specific and precise enough for consistency because of the direct financial consequences, but the texts clear enough for students to be willing to apply for aid.

We introduce 1,014 federal financial aid application instructional texts and identify jargon phrases using a glossary defined by the U.S. Department of Education. We find the text to be jargon-rich, with 40% of sentences including at least one jargon phrase and many sentences with multiple jargon phrases (see Figure 1). We show the results of a preliminary investigation in which we calculate the perplexity of a particular token using the cross-entropy loss measure by the unidirectional GPT-2 (Radford et al., 2019) language model head. We confirm the common-sense hypothesis that jargon tends to be more surprising than non-jargon and find there is a statistically significant difference between tokens within a multi-word jargon phrase.

*If you are eligible for a **Direct Loan** you need to complete a **Master Promissory Note**.*

Figure 1: A sentence from the corpus containing two jargon phrases, indicated in bold.

Data We introduce 1,014 federal financial aid application instructional texts from the official websites of a random sample of 341 Title IV U.S. higher education institutions[2]. The team manually curated these texts over the course of three consecutive application seasons; the texts were gathered in the fall of 2017, 2018, and 2019. Though the FAFSA application process has remained largely the same since 2015, there was a marked difference between 2018 and the other two years (See Table 1). This drastic change is most likely due to direction from President Barack Obama, who provided updated guidance in 2015 which did not take effect until the 2017-2018 FAFSA (Hoyt, 2015). These changes gave prospective college students the option to start the FAFSA earlier but also required earlier tax and income information, possibly complicating the application process, and thus, application instructions. To determine which phrases we annotated as jargon phrases, we used a glossary published by the U.S. Department of Education's Federal Student Aid Office. We identified 64 unique jargon phrases that appear at least 20 times

[1] We use *jargon* rather than 'terminology' to follow the tradition of prior research in education.

[2] A "Title IV institution" refers to any American post-secondary education that qualifies for federal financial aid.

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 380-383.
Held on-line February 14-19, 2021

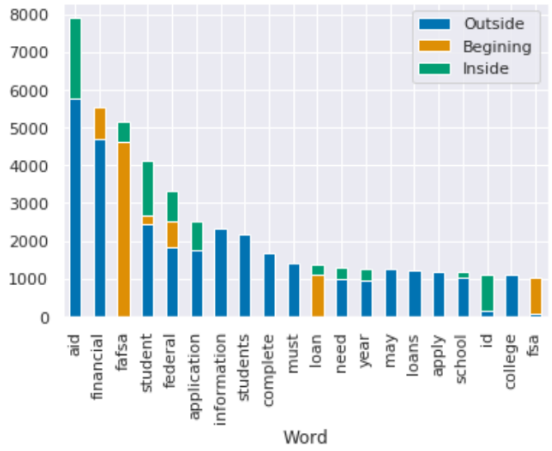

Figure 2: Distribution of 20 most common tokens. Tokens have been normalized and may not be exactly as they appeared in the text.

	Document Level			Sentence Level	
	All Tokens	Jargon Tokens	Sents	All Tokens	Jargon Tokens
2017	381.2	30.0	26.1	14.59	1.15
2018	669.0	45.7	47.6	14.05	0.96
2019	355.5	29.1	23.9	14.84	1.21
Average	250.5	35.8	32.6	14.4	1.07

Table 1: Average token and sentence distribution across the three years. For jargon phrases with more than one token, each token is counted individually (e.g. *Student Aid Report* is 3 tokens).

in dataset. The set of jargon phrases mostly consisted of noun phrases and included single-word terms such as *loan* as well as multi-word phrases such as *Promissory Note* and *Student Aid Report (SAR)*.

We annotated jargon phrases using the IOB scheme (Ramshaw and Marcus, 1999), where a token is either **I**nside, **O**utside, or **B**egins a particular jargon phrase. Each B-token was further annotated to indicate the jargon phrase id and type; if the jargon phrase had an identified acronym, we indicated whether the instance was the term with the acronym, just the term, or just the acronym. We identified a total of 7,616 jargon phrases consisting of 31,800 tokens. There is an average of 14.4 tokens per sentence, and 40% of the sentences contain at least one jargon phrase. Many tokens are found in more than one position; for example, the token *student* appears all three IOB positions: in *Student Aid Report (SAR)*, *Dependent Student*, and outside of a jargon phrase. We find that the jargon phrases are heavily skewed, with FAFSA and its variations accounting for nearly 70% of the identified jargon phrases.

Methods At a high level, we examine the 'fit' of jargon words in its context via perplexity from a large-scale pre-trained language model, GPT-2 (Radford et al., 2019). As the name suggests, perplexity is a quantitative measurement of how 'perplexed' the language model is when confronted with text; formally, it is defined as the exponentiated average log-likelihood of a sequence of words:

$$\begin{aligned} PPL(W) &= 2^{-\frac{1}{n}\log_2 P(w_1,\ldots,w_n)} \\ &= 2^{-\frac{1}{n}\sum_{i=1}^{n}\log_2 P(w_i|w_{<i})} \end{aligned} \quad (1)$$

Note that $-\log_2 P(w_i|w_{<i})$ quantifies the *surprisal* of the word w_i given its context $w_{<i}$ (Hale, 2001; Levy, 2008) following the seminal work of Shannon (1948), who suggested that when a word is likely to occur given a context, it communicates less information, thus taking less time to process.

Since GPT-2 is trained using a large amount of general-domain English, examining the perplexity differences in jargon and non-jargon tokens would allow us to understand how *surprising* tokens in a phrase are given their preceding contexts. This distributional analysis uses t-tests, where t-values represents a potential difference in the mean between two independent distributions. Here, we compare the perplexity distributions for tokens in different positions within a jargon phrase; comparing the perplexity of tokens that appear in the beginning of a jargon phase with tokens that appear outside of a jargon phrase.

To obtain the perplexity measurements, we make use of the cross-entropy loss of GPT-2 since cross entropy is the power term of perplexity. We run each sentence through the HuggingFace (Wolf et al., 2019) pre-trained implementation of the `GPT2LMHeadModel`.[3] In each experiment, we considered three measures of perplexity: **C**urrent

[3]GPT-2 uses a sub-word tokenizer; the results presented here measure the loss before and after the fist sub-word token of any word-level token as parsed by NLTK.

	Micro Average			Macro Average		
	O vs. B (df=386272)	B vs. I (df=32971)	I vs. O (df=384044)	OB (df = 1286)	B vs. I (df = 92)	I vs. O (df=1285)
Last	t = -7.236*	t = -8.497*	t = 17.807*	t = -1.046	t = -2.178	t = 4.815*
Current	t = -12.310*	t = 1.199*	t = 12.782*	t = -0.955	t = -1.389	t = 3.703*
Ratio	t = 8.912*	t = -32.541*	t = 3.358*	t = 0.384	t = -2.633	t = 3.965*

Table 2: Results for all t-tests. An asterisk $p < .0028$.

perplexity (mean after the first sub-word token of the current token), **L**ast perplexity (mean before the current token), and the **R**atio ($current/last$, representing a transition).

We completed t-tests at both a macro and micro level. We started with a micro-level analysis of the feature means, where each instance of each token carried the same weight in the compared distributions. There were no frequency minimums for this version. The distributions included 1,806 tokens instance at the beginning of a jargon phrase (B), 1,580 instances inside a jargon phrase (I), and 39,105 instances outside of a jargon phrase (O). For a macro-level analysis of the feature distributions, the mean perplexity was calculated for each token in each position before completing a t-test. Only tokens with a minimum frequency of 20 in a particular position were utilized in the test. The decision to require a minimum frequency of 20 meant there were 43 unique tokens that appeared at the beginning of a jargon phrase (B), 45 unique tokens that appeared inside a jargon phrase (I), and 1,038 tokens that appeared outside of a jargon phrase (O).

Findings The results of the completed distributional analysis, shown in Table 2, provide evidence that there may be some significant differences between tokens which appear in the IOB positions with regards to jargon phrases.

The micro-analysis gave the most evidence for differences between each position permutation. In this analysis, each of the position comparisons except one proved statistically significant for each perplexity feature, with a Bonferroni-corrected significance at $p < .0028$. The macro-analysis, which normalized the distributions taking the mathematical mean for a given token in a given position, suggests that only the inside and outside positions have a statistically significant difference in perplexity distribution.

In addition to the significance, we can also consider the direction of the t-values, which indicate whether a mean of a distribution was higher or lower than another. In both analyses, we can see that perplexity tends to be higher for tokens inside a jargon phrase than either token at the beginning of a jargon phrase or outside of a jargon phrase. These suggest that, the tokens that appear within jargon encode the most unexpected information, followed by tokens that appear at the beginning of a jargon phrase, followed by tokens that appear outside of jargon phrases.

Discussions This abstract presents a novel dataset and reveals the intuitive yet encouraging finding that jargon phrases trigger higher perplexity values in large-scale language models without any fine-tuning. This is encouraging, in the sense that most domain-specific texts that do not assume a barrier of entry also do not tend to come with a neatly compiled glossary; future work of jargon *discovery* could be informed by large general scale language models to improve probability-based models (Meyers et al., 2018). To this end, we attempted a prediction task using perplexity as the sole type of feature. We found, however, perplexity features alone were not able to reliably predict the presence of a jargon phrase, pointing to future work that seeks a deeper understanding of the semantics of jargon. Future lines of analysis also include comparing perplexity before and after fine-tuning to explore how perplexity of domain-specific phrases reacts to fine tuning.

Acknowledgements

We thank Katrin Erk and May Helena Plumb for reviewing early drafts of this paper, and all the members of the UT Austin Computational Linguistics group for valuable feedback and discussions. We also acknowledge the Texas Advanced Computing Center (TACC) at UT Austin for providing the computational resources for many of the results within this paper. Finally, we would like to thank the reviewers for their insightful and actionable feedback.

References

Mary Sonja Ardoin. 2013. Learning a different language: Rural students' comprehension of college

knowledge and university jargon.

Susan M Dynarski and Judith E Scott-Clayton. 2008. Complexity and targeting in federal student aid: A quantitative analysis. *Tax policy and the economy*, 22(1):109–150.

Mary Feeney and John Heroff. 2013. Barriers to need-based financial aid: Predictors of timely fafsa completion among low-income students. *Journal of Student Financial Aid*, 43(2):2.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Elizabeth Hoyt. 2015. President Obama Announces Changes to 2017-18 FAFSA.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2018. Attentionmesh: simple, effective and interpretable automatic mesh indexer. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 47–56.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Ismail El Maarouf, Mansar Youness, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. 2020. The finsim 2020 shared task: Learning semantic representations for the financial domain. *In Proceedings of IJCAIPRICAI 2020, Kyoto, Japan*.

Adam L Meyers, Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman, and Olga Babko-Malaya. 2018. The termolator: terminology recognition based on chunking, statistical and search-based scores. *Frontiers in Research Metrics and Analytics*, 3:19.

Julián Moreno-Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodriguez-Doncel, Artem Revenko, Sotirios Karampatakis, Maria Khvalchik, Christian Sageder, Jorge Gracia, and Filippo Maganza. 2020. Orchestrating nlp services for the legal domain. *arXiv preprint arXiv:2003.12900*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Kelly Ochs Rosinger. 2019. Can simplifying financial aid offers impact college enrollment and borrowing? experimental and quasi-experimental evidence. *Education Finance and Policy*, 14(4):601–626.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Zachary Wayne Taylor. 2020. College admissions for L2 students: Comparing L1 and L2 readability of admissions materials for us higher education. *Journal of College Access (2020)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.