GENTLE: A Genre-Diverse Multilayer Challenge Set for English NLP and Linguistic Evaluation

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, Amir Zeldes
Corpling Lab
Georgetown University
{ta571, sb1796, lg876, lel76, yl1290, yl879, sp1184, yz565, az364}@georgetown.edu

Abstract

We present GENTLE, a new mixed-genre English challenge corpus totaling 17K tokens and consisting of 8 unusual text types for out-of-domain evaluation: dictionary entries, esports commentaries, legal documents, medical notes, poetry, mathematical proofs, syllabuses, and threat letters. GENTLE is manually annotated for a variety of popular NLP tasks, including syntactic dependency parsing, entity recognition, coreference resolution, and discourse parsing. We evaluate state-of-the-art NLP systems on GENTLE and find severe degradation for at least some genres in their performance on all tasks, which indicates GENTLE's utility as an evaluation dataset for NLP systems.

1 Introduction

In the past several years, there have been great advances in NLP system performance on various tasks. However, many of these tasks are still evaluated on in-domain data, i.e. held-out data taken from the same domain as the system's training data. While this methodology is sound, it often overstates systems' ability to perform in real-world settings, where out-of-domain (OOD) data can lead to significant degradation (Plank, 2016; Joshi et al., 2018), even when target data comes from a similar domain (Nayak et al., 2020). For this reason, it is essential to have evaluation datasets with diverse text types, which can give a more accurate picture of systems' capabilities on OOD data, especially for domains that are distant from commonly studied domains or underrepresented in existing training datasets.

In this paper, we present GENTLE (**GEN**re **T**ests for **L**inguistic **E**valuation), a small but "extreme" open-access dataset that can be used for OOD evaluation of popular NLP tasks in English, as well as for linguistic analysis of less studied genres. The NLP tasks considered here include morphosyntactic tagging and dependency parsing according to Universal Dependencies (UD, de Marneffe et al. 2021), nested named and non-named entity recognition (NNER), coreference resolution, entity linking (Wikification), and hierarchical discourse parsing in the framework of Rhetorical Structure Theory (RST, Mann and Thompson 1988).[1] Our data comes from eight genres explicitly selected to represent unusual and diverse data types not currently included in the English Universal Dependencies corpora: dictionary entries, transcripts of live esports commentary, legal documents, medical notes, poetry, mathematical proofs, course syllabuses, and threat letters.

GENTLE enables us to answer various questions, including how well state-of-the-art (SOTA) models can parse OOD data and whether or not OOD genres are equally difficult for all NLP tasks. Apart from NLP performance, we can also see whether the annotation tasks in our challenge genres are difficult for humans and how the difficulties that arise in individual genres in GENTLE differ from those in existing datasets.

The rest of this paper is structured as follows: Section 2 presents some related work on OOD testing. Section 3 presents an overview of the corpus, while Section 4 compares the genres in the corpus in detail. Section 5 evaluates human agreement and NLP system performance on our data for each task, compared to more standard UD English data. Section 6 offers our conclusions. Our corpus is available at `https://github.com/gucorpling/gentle`.

2 Related Work

Previous work has focused on the importance of genre diversity and OOD evaluation for many of the NLP tasks included in GENTLE, supporting the general conclusion that NLP system performance tends to degrade on OOD data.

In coreference resolution, Moosavi and Strube

[1] The corpus is also openly released as part of the Universal Dependencies 2.12 version available at `https://github.com/UniversalDependencies/UD_English-GUM`.

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 166–178
July 13, 2023

(2017) and Zhu et al. (2021) point out that existing models mainly rely on lexical features (e.g. word embeddings) and may face the problem of overfitting because of the large overlap of vocabulary between training and testing data. Apart from overfitting, low recall resulting from domain discrepancy is another major problem for named entity recognition (NER, Augenstein et al. 2017).

Despite a recent surge in approaches for discourse-level tasks, there is still room for improvement in this area, especially for OOD data (Atwell et al., 2021). Liu and Zeldes (2023) investigate the impact of genre diversity in training data composition for RST discourse parsing, the task of recursively identifying relations between propositions. They show that diverse data is essential for stable and generalizable models for this task.

Similarly to the present work, Kanerva and Ginter (2022) conduct an OOD evaluation of Finnish dependency parsing, including constructing a relatively "extreme" OOD treebank, including 5 distinct genres (web documents, clinical, online discussions, tweets, and poetry).[2] Their experiments indicate that syntactic parsing performance degrades severely on OOD data, particularly on the LAS (labeled attachment score) metric.

Data diversity is thus crucial for a range of NLP tasks, but the lack of diverse data available hampers training and evaluation. Previous corpus construction efforts cover a wide range of English genres, for example, 5 genres in the English Web Treebank (EWT, Silveira et al. 2014) for syntactic annotations, and 6 in OntoNotes (Weischedel et al., 2012) for NER and coreference as well. However, both datasets lack nested, non-named entities, entity linking (Wikification), and discourse parsing.

More recently, the UD English GUM corpus (Georgetown University Multilayer corpus, Zeldes 2017), with data from 12 genres (academic articles, biographies, conversation transcripts, works of fiction, Reddit posts, how-to-guides, interviews, news articles, political speeches, textbook excerpts, Wikivoyage travel guides, and YouTube vlog transcripts), covers all of the annotations examined in this paper, and raises the expectation of being a possibly good training set for OOD targets, due to its diverse content. Experiments in this paper will therefore use our newly annotated OOD GENTLE corpus to evaluate SOTA models trained on the already diverse GUM corpus and compare their performance on both datasets.

3 GENTLE

The GENTLE corpus is constructed as an OOD evaluation dataset, modeled on the test set for the English GUM corpus. Table 1 gives an overview of partitions in GUM (v9.0) compared to GENTLE.

dataset	genres	docs	tokens
$\text{GUM}_{\text{train}}$	12	165	160,700
GUM_{dev}	12	24	21,409
GUM_{test}	12	24	21,770
GENTLE	8	26	17,797

Table 1: GUM Partitions vs. GENTLE.

GENTLE forms an extension to the GUM test set with 8 more genres, for a total of 20 diverse text types to test on. Although the amount of data in GENTLE is small, the data follows GUM's scheme and is richly annotated on many layers, containing over 250K key-value annotations connected by complex annotation graphs. For treebanking, the annotations include gold-standard layers for Universal Dependencies morphosyntax, such as XPOS (Penn Treebank) tags, lemmas, and basic dependencies. In addition, automatically-derived morphological features, enhanced dependencies and UPOS tags are obtained using the DepEdit library (Peng and Zeldes, 2018) with the same scripts that produce these layers for the GUM corpus.

For NNER and coreference resolution, the data includes nested, named and non-named entity annotations. These employ the same scheme used in GUM, with 10 entity types, 6-way information status annotations, coreference and bridging links (9 edge types from GUM, including split antecedents, discourse deixis, etc., see https://gucorpling.org/gum/). GUM-style entity linking (wikification, Lin and Zeldes 2021) is also provided, with an automatically produced alternate version of the entity/coreference annotations matching the OntoNotes scheme (Weischedel et al. 2012; see Zhu et al. 2021 for details). The data also includes complete hierarchical discourse trees in Rhetorical Structure Theory (RST, Mann and Thompson 1988), following the same scheme as GUM.

Annotation was conducted by the authors of this paper during several hackathon-style annotation sessions. Although varying in expertise on each task, every annotator had previous experience annotating every layer of annotation described

[2] https://github.com/UniversalDependencies/UD_Finnish-OOD/

above. For annotation tools, morphosyntactic layers (XPOS tags, lemmas, and basic dependencies), entity layers (entity and coreference), and discourse layers (EDU segmentation and discourse relation) were annotated on Arborator (Gerdes, 2013) and Midas Loop (Gessler et al., 2022), GitDox (Zhang and Zeldes, 2017), and rstWeb (Zeldes, 2016), respectively. We also double annotated a portion of the corpus to measure human agreement, which will be further described in §5.

In choosing data, we attempted to select challenging types of spoken and written open-access materials that are maximally different from those already found in GUM (cf. §2). Texts were selected for each genre from a single source, making sure that (1) the total number of tokens falls between 2k and 2.5k tokens per genre; (2) at least 2 texts are selected to better represent the genre (as mean can only be calculated with 2 or more documents per genre). While the texts were selected randomly for most genres, the texts for some genres were manually selected. For instance, since poetry texts can be extremely short, the documents for this genre were chosen to be varied in length, as to limit the number of documents needed to reach the target token range. Table 2 gives the genre composition and sources for each data type in GENTLE.[3]

genre	docs	tokens	source
`dictionary`	3	2,423	Wiktionary
`esports`	2	2,149	YouTube
`legal`	2	2,288	Wikisource / CUAD[4]
`medical`	4	2,164	MTSamples
`poetry`	5	2,090	Wikisource
`proof`	3	2,106	Proofwiki
`syllabus`	2	2,431	GitHub
`threat`	5	2,146	casetext
total	26	17,797	

Table 2: Corpus Contents of GENTLE.

The chosen data is broad not only in domain, including medical, legal, and other technical areas, but also in medium (online linked resources such as Wiktionary data, spontaneous spoken esports commentary, and threat letters) and communicative intent (e.g. poetry, syllabuses, and mathematical proofs). These genres can also be challenging for both humans and NLP models, as they diverge in various ways from standard training data and materials that guidelines are based on for each task.

[3]Please refer to Appendix A for detailed information on the contents of each genre.

[4]The **C**ontract **U**nderstanding **A**tticus **D**ataset (CUAD) v1 from the The Atticus Project (Hendrycks et al., 2021): `https://www.atticusprojectai.org/`

genre	slen	pass	n/v	ttr	oov	sglt
GUM	20.16	.07	2.36	.4	–	.29
GUM$_{news}$	22.52	.12	3.34	.45	–	.28
dictionary	10.98	.1	3.65	.39	.11	**.49**
esports	21.07	**.01**	1.48	.36	.08	.24
legal	21.58	.04	3.33	.36	.17	.33
medical	11.21	.15	4.31	.46	.22	.32
poetry	17.7	**.01**	1.59	**.53**	.11	.25
proof	15.63	**.18**	5.14	**.25**	**.24**	**.13**
syllabus	**7.65**	.12	**5.34**	.43	**.24**	.38
threat	**24.25**	.02	**1.3**	.49	**.05**	.28

Table 3: Average sentence length (`slen`), passive ratio (`pass`), noun/verb ratio (`n/v`), type-token ratio (`ttr`), out-of-vocabulary ratio (`oov`), and singleton ratio (`sglt`).

Before approaching a technical evaluation of how well humans can annotate these materials (inter-annotator agreement) and how NLP models score on them for each task, in the next section, we explore how the materials differ from genres in GUM descriptively, in text content and annotations.

4 Variation across Genres

4.1 Summary Statistics

Because the materials in GUM and GENTLE cover a vast range of text types, a quantitative view of variation in the data can provide a useful starting point in understanding what makes each genre unique. Although we could also devote as much attention to GUM genres, for space reasons, we will focus here on how each GENTLE genre is distinct from GUM and other genres (for more on GUM genres, see Zeldes and Simonson 2016).

Table 3 gives an overview of some commonly used descriptive metrics to compare GENTLE genres to the GUM corpus average, as well as the score for GUM's `news` genre, which can be taken as a stand-in for the standard language typically found in reference corpora, e.g. the Wall Street Journal (Marcus et al., 1993). The lowest and highest numbers in each metric are colored in red and blue.

Most genres in GENTLE have substantially shorter sentences (**slen**) than the GUM average, with `syllabus` having the lowest mean of 7.65 tokens, largely due to frequent bulleted or numbered lists of course topics, which are noun phrase fragments (e.g. *Week 3 - JavaScript Fundamentals*). The only genre with substantially above average sentence length is `threat`, in which long and sometimes rambling justifications or elaborate

	UPOS		dependency relations		entity types		discourse relations	
GUM	PROPN	↓-25.13	dep	↓-20.93	org.	↓-19.50	joint	↓-8.89
GUM$_{news}$	PROPN	↑41.41	flat	↑21.74	org.	↑30.60	attrib.	↑12.14
dictionary	X	↑21.16	punct	↑20.37	abstract	↑17.05	org.	↑5.72
esports	ADV	↑5.91	parataxis	↑16.68	event	↑9.94	eval.	↑6.82
legal	X	↑18.37	dep	↑17.20	org.	↑12.62	context	↓-3.69
medical	NOUN	↑11.59	nummod	↑7.88	substance	↑11.02	joint	↑12.86
poetry	PROPN	↓-7.07	compound	↓-5.68	animal	↑17.84	mode	↑5.50
proof	SYM	↑56.27	dep	↑10.92	abstract	↑39.23	explan.	↑8.46
syllabus	X	↑54.48	dep	↑46.31	abstract	↑25.69	joint	↑18.23
threat	PRON	↑13.17	punct	↓-6.98	person	↑12.91	explan.	↑6.43

Table 4: Strongest Standardized χ^2 Residual Label in 4 Layers for each Genre.

consequences are often added to main sentences.

Passivization (**pass**) is rare overall, except for medical texts (double the GUM average) and math proofs (even more), in which volitional agents are often suppressed (in the former, someone was *diagnosed* but we do not know by whom; in the latter, a variable *can be assigned*, etc.).

Noun/verb ratio (**n/v**) and type-token ratio (**ttr**) reveal that `syllabus` has a rich and mainly nominal vocabulary (lists of skills or topics, primarily nouns/compounds). Though rich in **ttr**, `threat` is more verbal. `poetry` has the highest **ttr**, partly because some poetic constraints discourage repetition (e.g. alliteration and rhyming, where duplication is avoided). In contrast, `proof` has the lowest **ttr** since some terms are used repeatedly (e.g., *vertex* is repeated ten times in one proof about vertices).

The out-of-vocabulary (**oov**) rate shows the percentage of tokens in each genre that is not attested in GUM, which can be expected to correlate with NLP tool degradation. `proof`, `syllabus` and `medical` have extremely high rates (nearly 25% of tokens are never seen in GUM), while `threat` and `esports` have less alarming rates of 5–8%.

Finally, the proportion of singleton mentions (**sglt**, entities referred to just once in a text) shows that `proof` documents have repetitive vocabularies and repeatedly refer to the same entity. This is because once a member or a class of possible items has been introduced, its properties are discussed in detail (e.g., after defining *Let DE be a rational straight line*, we may continue discussing the line DE). By contrast, `dictionary` documents use many arbitrary entities in example sentences that are never mentioned again (in an example sentence for *school*, we find *Harvard University is a famous American post-secondary school*, but *Harvard* is then never mentioned again). These genre disparities and unique environments can be expected to interfere with prior probabilities learned by NLP models, and, as we will see below, also with human annotation agreement.

4.2 Label Distributions

To give a quick overview of which labels deviate from their expected frequency in each genre, Table 4 gives standardized chi-square residuals in a contingency table of labels versus genres. A positive residual means that a label is used more frequently than expected based on its overall frequency, and a negative residual means the opposite – that a label is used less frequently than expected. Here we give only the strongest deviation associated with each genre in each of four annotation layers (for the complete tables of residuals, see Appendix B).

The deviation with the absolute highest score in the parts-of-speech (UPOS) is the unsurprising frequency of the tag `SYM` in math proofs, used for many mathematical symbols. The second highest is the tag `X` in `syllabus`, used to tag bullet point markers and also used frequently in `legal` documents. Other tag deviations include the lack of proper nouns in `poetry`, dense use of punctuation in dictionary entries, and the prevalence of common nouns in the medical data (a lack of pronouns mirrors this, see Table 7 in Appendix B).

Dependencies show some parallel phenomena (`punct` in `dictionary`, `dep` in `legal` and `syllabus`, which is used to attach bullet points), but also reveal lack of punctuation in threat letters. The prevalence of `parataxis` in `esports` to narrate chains of events as they unfold is also noteworthy, as in (1), and the use of numerical quantities in medical texts, often used for medication dosages as in (2). The `poetry` genre shows a negative deviation in avoiding nominal compounds, which are more typically a property of technical texts in English, e.g. in nested noun-noun compounds found in medical notes, as in (3).

(1) *Jović scoring, van de Beek and Ibrahimovic coming on 3-1 ...*

(2) *Prilosec 20 mg b.i.d.*

(3) *white blood cell count*

Residuals of entity types also expose differences compared to GUM genres and news in particular, which distinguishes itself by frequently mentioning organization entities. proof is the most extreme in favoring the abstract type (in fact, over 96% of mentions in proof are abstract), while threat focuses on people. medical is unique with its preponderance of substance entities, primarily medications, while esports disproportionately uses the event type. One result in the table is an artifact of one specific document, and the small corpus size: animal in poetry is due entirely to the inclusion of Edgar Alan Poe's "The Raven".

Finally, discourse relations reveal the prevalence of coordinated lists annotated in the relation class JOINT in syllabus (topics, assignments, weeks in the course, etc.) and medical (symptoms, vital statistics, medications; all mainly the relation subtype JOINT-LIST); esports unsurprisingly favors EVALUATION to convey positive or negative impressions of players, and poetry is unique in favoring MODE relations, primarily due to the relation subtype MODE-MANNER, which is used in adverbial manner adjuncts or parataxis, as in (4)–(5). legal shows a negative tendency to avoid CONTEXT relations, which include background and spatio-temporal contextual information, both of which are less needed in a highly specialized and professional text in which context is often a given and statements apply in general.

(4) *I sat divining, [with my head at ease]*$_{\text{MANNER}}$

(5) *[We slowly drove]*$_{\text{MANNER}}$ *He knew no haste*

4.3 Proximity across Genres

The metrics in §4.2 reveal differences among GENTLE genres compared to GUM. But needless to say, there are also many similarities between the GENTLE and GUM genres. To describe proximity across genres, we utilize the features in Table 3 and the full residual tables for the four annotation layers in Appendix B to build a cluster dendrogram of GENTLE and GUM genres.

Because labels occupy different numerical ranges and have diverse tag set sizes (only 10 entity types but 34 coarse dependency labels), we scale the data by transforming it into z-scores, and then reduce the dimensionality of each table of residuals to five columns using Principal Component Analysis (PCA). In other words, while the original table of entity residuals has one row per genre and ten columns for the entity types (Table 9 in the Appendix) and contains chi-square residuals, the transformed table is based on a z-scaled version of the same table, which is reduced to having only 5 total columns using PCA. This affords each annotation layer as much space as the five features in Table 3 (excluding OOV rate, which is inapplicable to GUM data), for a total of 25 features per genre (the five scaled metrics without OOV, and five features each for POS, dependencies, entities, and discourse relations).

Because we are interested in concord/discord between genres across layers and do not necessarily care if z-scores are more or less extreme for a particular annotation layer, we use ordinal Kendall correlations between values of each dimension to compute the distance metric between genres, thereby avoiding single features with large values dominating the clustering. In the ordinal clustering, genres are closer if their ranks for multiple features are ordered more similarly—e.g., if they are ranked first and second in type-token ratio and singletons, then those two genres display positive concord along those features. We apply single linkage clustering to produce the dendrogram in Figure 1.[5]

As the figure shows, several of the GENTLE genres (in red) form outliers and cluster apart from genres in GUM (in blue). This suggests, on the one hand, they are substantially distinct and, therefore, valuable additions to already available genres in GUM. On the other hand, they may be challenging to handle for models trained on GUM. This is especially true for genres like proof on the left side of the plot, which forms the most distinct outlier, in a top-level cluster of its own, and quite distant vertically from other genres. We can also see legal quite distant from its nearest neighbors,

[5] An anonymous reviewer has inquired whether we attempted other clustering procedures: the answer is yes—the decision to use ordinal clustering resulted from the observation that single annotation layers had outsize influence for some genres, such as SYM tags in proof; single linkage is both a default choice, and works well to cluster pairs of near genres as dendrogram leaves.

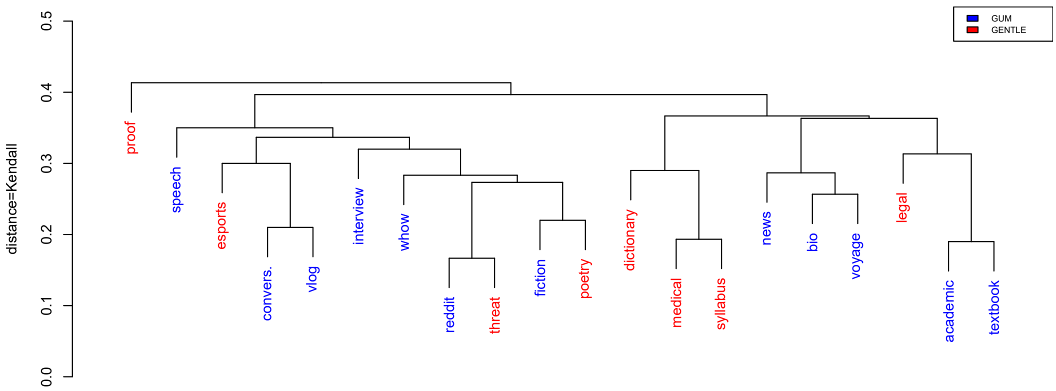

Figure 1: Cluster Dendrogram for GUM and GENTLE Genres.

GUM's `academic` and `textbook`, which are near each other. Three GENTLE genres, `dictionary`, `medical`, and `syllabus` form a sub-cluster, with the latter two being relatively similar, possibly due to both genres being dominated by bulleted lists comprised of noun phrases, i.e., sentences fragments.

In the middle, `poetry` is the closest to GUM's `fiction`, perhaps partly due to long sentences, extensive vocabulary, and verb-dominated morphosyntax. `threat` clusters with GUM's `reddit` genre, perhaps because both are relatively argumentative genres, often written in first-person, and include many interjections and swearwords (see Behzad and Zeldes 2020 for similar and additional observations on Reddit data). `esports` is somewhat far from its nearest neighbors, the informal spoken genres `conversation` and `vlog` which intuitively share features and cluster together; the latter also comes from the same source and modality as `esports`, since both were collected from YouTube. GUM's more informative expository genres also cluster together plausibly, with biographies (`bio`) and travel guides (`voyage`) grouped together after the split with `news`.

5 Evaluation

To understand how challenging GENTLE data is for both NLP models and humans, we evaluate representative systems on each task using the entire corpus and conduct an inter-annotator agreement (IAA) experiment by double annotating 10% of the data. Table 5 reports Cohen's Kappa (κ) and task-specific scores where applicable, taking the gold standard release data as a reference, compared to a second human's annotation. The double annotations were done without additional validation checks; in other words, the final gold data, subjected to stringent validations by the official UD validator and validation scripts from the English GUM repository, can be expected to be more consistent and reliable. Double annotated data comes from document initial "snippets" in each genre since non-initial sections may be incoherent for layers such as coreference. Each snippet was around 200-250 tokens in length, amounting to 1,838 tokens in total ($\approx$10.34% of the entire corpus).

However, it is also true that NLP accuracy in document-initial positions diverges from overall accuracy since documents are systematically non-homogeneous. For example, `dictionary` entry beginnings are much harder to parse since they contain technical notation, foreign language etymologies, and more, while later sections typically include grammatically simple usage example sentences. Therefore, we report NLP accuracy on the double annotated snippets compared to human scores in Table 5, separately from the overall performances on the GENTLE corpus in Table 6. For each setting, we report scores by genre, for the entire corpus (micro-average), and the averaged per-genre score (macro-average). All NLP models were trained on the GUM v9 train partition and tested on the established GUM v9 test set and GENTLE. Additionally, we include genre-specific numbers for GUM's `news` section, which can be taken to represent the most commonly used evaluation data type in most NLP tasks.

Tokenization, Tagging, Lemmatization, and Dependency Parsing We use the widely employed Stanza package (Qi et al., 2020) to evaluate gold-

Tasks	Metrics	MICRO	MACRO	`dictionary`	`esports`	`legal`	`medical`	`poetry`	`proof`	`syllabus`	`threat`
					Human Agreement on Snippets						
POS Tagging (XPOS)	Acc	95.38	95.37	94.69	**98.25**	**93.48**	94.81	97.85	95.67	93.86	94.37
	κ	94.98	94.78	95.38	**98.00**	93.46	94.49	97.29	94.38	**92.08**	92.94
Lemmatization	Acc	96.90	96.89	**92.92**	**99.56**	95.22	96.54	97.42	98.70	95.61	99.13
	κ	96.86	96.82	**92.65**	**99.55**	95.12	96.47	97.36	98.66	95.56	99.11
Dependency Parsing	UAS	88.79	88.77	**77.88**	85.53	90.00	88.74	90.13	88.74	93.86	**95.24**
	LAS	84.66	84.63	**73.01**	81.58	83.48	87.01	88.41	83.55	89.47	**90.48**
Entity Recognition (untyped)	P	89.47	89.25	93.24	92.54	91.94	79.71	**78.08**	**96.19**	86.36	95.71
	R	85.27	84.84	81.18	92.54	79.17	**77.46**	82.61	**97.12**	84.44	83.75
	F	87.32	86.88	86.79	92.54	85.07	**78.57**	80.28	**96.65**	85.39	89.33
Entity Recognition (typed)	P	81.91	81.35	90.54	**70.15**	90.32	73.91	76.71	**96.19**	73.86	78.57
	R	78.06	77.32	78.82	70.15	77.78	71.83	81.16	**97.12**	72.22	**68.75**
	F	79.94	79.19	84.28	**70.15**	83.58	72.86	78.87	**96.65**	73.03	73.33
Coreference Resolution	MUC	70.46	66.01	47.05	**94.44**	72.22	60.86	62.06	70.58	**38.09**	82.75
	B^3	77.63	77.21	83.50	**90.29**	75.31	65.65	**62.97**	84.74	76.38	78.87
	$CEAF_{\phi 4}$	72.25	70.55	84.43	**86.38**	69.30	63.55	**48.10**	74.50	73.46	64.70
	Avg. F	73.45	71.26	71.66	**90.37**	72.28	63.35	**57.71**	76.61	62.64	75.44
					NLP Performance on Snippets						
XPOS	Acc	92.56	92.55	86.73	97.66	95.36	97.55	**97.71**	**77.63**	93.27	94.52
Lemmatization	Acc	96.32	96.33	97.64	**99.56**	97.10	96.25	**92.56**	94.81	94.44	98.27
Dependency Parsing	UAS	80.69	80.65	65.34	85.23	87.83	87.01	**90.41**	**54.69**	85.09	89.61
	LAS	76.22	76.18	59.00	79.39	82.75	83.41	**87.55**	**50.65**	81.14	85.57
Entity Recognition (typed)	P	75.63	75.14	72.22	70.42	**66.89**	74.86	72.80	**84.91**	78.42	80.60
	R	70.01	69.81	**60.61**	64.88	61.72	71.21	69.80	73.33	71.40	**78.26**
	F	72.71	72.34	65.91	67.53	**64.20**	72.98	71.27	**82.67**	74.74	79.41
Coreference Resolution	MUC	65.66	54.86	**0.00**	**83.72**	30.30	80.95	74.62	52.30	42.85	74.15
	B^3	41.25	36.72	**4.49**	54.27	22.78	38.73	56.33	26.45	29.47	**61.23**
	$CEAF_{\phi 4}$	17.72	18.31	**1.80**	22.00	20.95	6.82	**36.13**	14.32	15.79	28.67
	Avg. F	41.54	36.63	**2.10**	53.33	24.68	42.17	**55.69**	31.02	29.37	54.68

Table 5: Human Performance and Corresponding NLP Performance on GENTLE Snippets for 5 NLP Tasks. The highest scoring ('easiest') GENTLE genres are highlighted in **blue**, and the lowest scoring are in **red**.

tokenized texts in Table 5 allowing comparisons with human agreements, as well as end-to-end from plain text in Table 6 to also evaluate tokenization. Tokenization degrades in the end-to-end scenario for all GENTLE genres except for `threat`. Tokenization is error-prone in `syllabus` and `legal` due to the abundance of bulleted and numbered nominal phrases and abbreviations. XPOS tagging degrades nearly 10 points on GENTLE and scores the lowest on `proof` and `syllabus` due to mathematical symbols (e.g. $\leqq$, $\in$, x, y) and genre-specific terminologies (e.g. TAs, TBD). Micro-averaged lemmatization performance drops nearly 6 points to 92.38 and parsing by 15 points to a LAS of 72.38, again worst for `proof` and `syllabus`.

While these results may be somewhat shocking, human performance is also imperfect, with XPOS and lemmatization accuracy in the mid-90s, less than 3 points above Stanza for tagging, and neck-and-neck for lemmatization, and with human LAS at 84.66, about 8 points above Stanza on average. To illustrate why humans disagree on syntax especially in technical genres, we offer a brief example of parsing a `legal` case law designation for '410 U.S. 113' in Figure 2. '410 U.S.' is a volume of US Supreme Court cases, including case '113' (Roe v. Wade) – one annotator (in black) analyzes '113' (the case) as the head, which is modified by the name of the volume that includes it, while the other treats the volume as the head, with a numerical modifier attached as dep, similar to how GUM annotates cases like 'Page 5.' Without good intuitions about Supreme Court case nomenclature and very clear guidelines, any chance of perfect agreement is hampered by a myriad of such cases.

On the other hand, some potentially difficult genres, such as `esports`, turned out to have high human agreement for tokenization, tagging and lemmatization, despite well known challenges in annotating User Generated Content (UGC, see Sanguinetti et al. 2022).

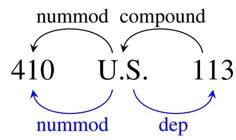

Figure 2: Annotation Disagreement for *410 U.S. 113*.

Tasks	Metrics	GUM_{test}	$\text{GUM}_{\text{test-news}}$	GENTLE (MICRO)	GENTLE (MACRO)	`dictionary`	`esports`	`legal`	`medical`	`poetry`	`proof`	`syllabus`	`threat`
Tokenization	F	**99.61**	99.67	97.29	97.46	98.12	99.52	95.55	97.73	99.59	97.98	**91.46**	**99.69**
POS Tagging (XPOS)	Acc	**97.46**	**97.85**	88.34	88.56	90.74	**95.89**	89.71	92.93	91.51	78.76	**75.22**	93.74
Lemmatization	Acc	**98.13**	**98.52**	92.38	92.64	95.53	**98.29**	91.72	93.01	95.51	91.06	**79.74**	96.23
Dependency Parsing	UAS	**89.49**	**89.68**	76.71	77.01	75.39	83.99	77.23	81.15	76.74	71.23	**63.99**	**86.37**
	LAS	**87.21**	**87.45**	72.38	72.65	70.78	78.84	73.95	77.64	71.70	65.58	**59.94**	**82.77**
Entity Recognition (typed)	P	**77.14**	65.01	75.63	75.10	72.22	70.42	**66.56**	74.86	72.80	**84.91**	78.42	80.60
	R	**76.24**	72.69	70.01	69.77	**60.61**	64.88	61.41	71.21	69.80	**80.56**	71.40	78.26
	F	**76.88**	68.64	72.71	72.30	65.91	67.53	**63.88**	72.98	71.27	**82.67**	74.74	79.41
Coreference Resolution	MUC	**76.38**	59.67	60.89	55.98	**9.30**	67.84	59.14	70.13	70.92	48.95	41.09	**80.48**
	B^3	**64.71**	53.97	33.37	33.91	**14.74**	45.49	31.07	32.78	43.98	29.08	20.88	**53.29**
	$\text{CEAF}_{\phi 4}$	**57.15**	**53.06**	9.75	11.18	**4.91**	**17.48**	9.22	7.06	15.10	13.48	7.78	14.38
	Avg. F	**66.08**	55.57	34.67	33.69	**9.65**	43.60	33.14	36.66	43.33	30.50	23.25	**49.38**
RST EDU Segmentation (Gold)	P	**96.43**	**95.68**	93.90	93.21	**97.58**	95.71	90.07	**97.58**	91.30	**88.81**	94.35	93.62
	R	**95.85**	**97.17**	93.17	92.07	95.48	**87.01**	96.11	96.58	88.06	87.89	**98.04**	92.69
	F	**96.14**	**96.42**	93.53	92.60	96.52	91.16	92.99	**97.07**	89.66	**88.35**	96.16	93.16
RST EDU Segmentation (Trankit)	P	**93.63**	**92.91**	89.90	90.17	95.48	94.37	85.29	**97.92**	87.46	87.41	**80.48**	92.98
	R	**93.48**	**96.46**	86.78	87.79	86.24	87.01	92.23	**96.58**	85.48	88.93	**73.48**	92.36
	F	**93.55**	**94.65**	88.31	88.89	90.62	90.54	88.62	**97.24**	86.46	88.16	**76.82**	92.67
RST Parsing	S	**70.07**	**71.89**	62.15	62.83	59.31	**55.77**	**72.72**	65.51	59.78	69.11	57.13	63.29
	N	**56.90**	**60.61**	47.63	48.05	47.47	**40.41**	**59.79**	50.35	40.87	55.25	44.18	46.06
	R	**49.57**	**56.40**	37.64	38.16	30.52	**29.30**	**51.48**	46.88	30.93	41.73	40.17	34.23

Table 6: End-to-End NLP Performance on All Tasks on Full Plain Texts (averaged over 3 runs). Top and bottom scoring GENTLE genres are marked in **blue** and **red** (GUM scores are nearly always higher, in **bold**).

Entity Recognition and Coreference Resolution For NNER, we evaluate a SOTA neural system (seq2set, Tan et al. 2021). In both full GENTLE and snippets, we consider plain text with gold tokenization as input and use precision, recall and F1 to evaluate. In Table 6, F1 drops over 4 points on average, and over 13 points on `legal`. Inspection reveals most errors involve malpredicted spans, especially when deciding entity boundaries with PP attachment, apposition, or coordination. For example, in *[Proto-Germanic *nēhwist (“[nearest]$_2$, [closest]$_3$”)]$_1$*, span 2 (blue) and span 3 (orange) are appositions providing additional information for the word *nēhwist* and span 1 (red) as a whole forms a non-named entity span, but neither of them are correctly predicted by the model. `proof` outperforms GUM because mathematical variables, which are frequent in `proof`, are easier to identify compared with other types of entities. We also observe this in Table 5, where IAA is the highest for `proof`. Note that IAA for typed and untyped entities are identical; this is because most entities in `proof`, e.g. mathematical variables, are *abstract*.

The coreference resolution task uses MTL-coref (*under review*), a new SOTA model for the GUM benchmark which is trained with singletons and other entity-level information. We use the F1-measure of MUC, B^3, $\text{CEAF}_{\phi 4}$, and the average CoNLL score as evaluation metrics. Table 6 reveals that the model performs substantially worse on GENTLE, with nearly 32 points degradation. Genre-wise analysis reveals that `dictionary`, which has few pronouns, performs worst, while `threat`, rich in pronouns, scores best in GENTLE. This shows that the model struggles more with complex NPs (with possible PP attachments) and proper nouns but can more easily identify coreference chains involving pronouns (and especially the easy pronouns ‘I’ and ‘you’ in `threat` letters). For instance, in `GENTLE_epsorts_fortnite`, the model incorrectly clusters *[Kreo]$_1$* ... *[him]$_1$* ... *[Maufin]$_1$*, a chain including multiple names unseen during training.

RST Segmentation and Parsing We evaluate GENTLE on two RST tasks: elementary discourse unit (EDU) segmentation and RST parsing. For EDU segmentation, we use DisCoDisCo (Gessler et al., 2021), the winning system in the 2021 DISRPT shared task on segmentation. We evaluate EDU segmentation under two conditions: ‘Gold’, where the full, human-provided UD parses for GENTLE documents are provided to the system; and ‘Trankit’, where with the sole exception of tokenization (which remains human-provided), all UD parse information is provided by Trankit’s (Nguyen et al., 2021) default English model.

For RST parsing, we use the best setting from the bottom-up neural parser by Guz and Carenini (2020), `SpanBERT-NoCoref`, which obtained the SOTA performance on GUM as of v8 (Liu and Zeldes, 2023) using the original Parseval procedure on binary trees, following Morey et al. (2017). We evaluate using gold discourse units for simplicity and comparability with previous work.

Unsurprisingly, GENTLE contains challenging materials even with gold discourse units: overall, the best-performing genre is `legal` while the worst-performing genre is `esports`. By examining dependency conversions of gold vs. predicted trees following Li et al. (2014), we found that the model was only able to correctly identify the Central Discourse Unit in 6 out of 26 documents (23.1%) in GENTLE. The top 2 most difficult relation classes are TOPIC and EXPLANATION, both of which tend to lack explicit and unambiguous cues such as discourse markers, and may require an understanding over multiple EDUs.

6 Conclusion

We have introduced GENTLE, a new, genre-diverse, richly-annotated test corpus for English. While this new resource is relatively small, the challenging genres included in the corpus are diverse not only in topic, but also in terms of medium and communicative intent. The 8 genres have considerably distinct characteristics reflected in metrics and label distributions for individual annotation layers. These genres also differ substantially from the 12 genres in the GUM reference corpus. As such, GENTLE serves as an important complement to GUM's test set, and can provide valuable insights into NLP systems' ability to perform on OOD data.

We found in evaluations that system performance generally degraded on GENTLE compared to GUM, corroborating prior findings that NLP systems degrade on OOD data. However, degradation was not uniform, and different genres presented differing degrees of difficulty for different NLP tasks. For dependency parsing, the steepest degradation was in `syllabus` and `proof`, while entity recognition saw particularly poor performance in `legal` and `dictionary`, and RST parsing performed lowest on `esports`, `dictionary` and `poetry`. It is thus necessary to have a wide variety of genres available for evaluation if one aims for a holistic understanding of the capabilities and limitations of an NLP system.

Moreover, it is worth noting that the annotation tasks for our challenge genres were not just difficult for the NLP systems, but for our human annotators as well. Our IAA experiments showed that human annotation generally outperformed the NLP systems in terms of accuracy. However, some genres stood out as being particularly difficult for humans, such as `dictionary`, which suggests that it would be beneficial to develop additional annotation guidelines targeting difficult cases that arise from genre-specific phenomena.

With the introduction of GENTLE and the results from the above evaluation experiments, we hope to encourage the use of genre-diverse test corpora for NLP benchmarks. This will allow researchers to obtain realistic measures of how NLP systems will perform on OOD data, which is frequently the use case of interest in real-world applications of NLP technologies.

Limitations

Our corpus is designed to serve as a challenge set, and is limited in size: each of the 8 genres ranges from 2k to 2.5k tokens, totaling around 18k tokens. Given the amount of work necessary for multilayer annotations, building a larger challenge set was not deemed realistic with the limited resources available for this project, and is left for future work.

Additionally, the evaluation of inter-annotator agreement is limited to a small amount of data, since double annotating the amount of annotation layers involved is costly. In particular, the evaluation is limited by the use of a common gold tokenization standard to facilitate reporting commonly used scores (Cohen's Kappa, tagging accuracy, NNER F1, etc.), which do not reflect cascading errors due to tokenization disagreements. Additionally, we did not perform double annotation experiments for RST discourse parsing, as these would require annotating entire documents in each genre, which would exceed the amount of data we were able to have annotated for this evaluation.

References

Mitchell Abrams. 2019. Uncovering the genre of threatening texts: A multilayered corpus study. Master's thesis, Georgetown University.

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. Association for Computational Linguistics.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Shabnam Behzad and Amir Zeldes. 2020. A cross-genre ensemble approach to robust Reddit part of speech tagging. In *Proceedings of the 12th Web as Corpus Workshop*, pages 50–56, Marseille, France. European Language Resources Association.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kim Gerdes. 2013. Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 88–97, Prague, Czech Republic. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luke Gessler, Lauren Levine, and Amir Zeldes. 2022. Midas loop: A prioritized human-in-the-loop annotation for large scale multilayer data. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 103–110, Marseille, France. European Language Resources Association.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *NeurIPS*.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1199, Melbourne, Australia. Association for Computational Linguistics.

Jenna Kanerva and Filip Ginter. 2022. Out-of-domain evaluation of Finnish dependency parsing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1114–1124, Marseille, France. European Language Resources Association.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2021. WikiGUM: Exhaustive entity linking for wikification in 12 genres. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 170–175, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? A thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3104–3122, Dubrovnik, Croatia. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics*, 19(2):313–330.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Siyao Peng and Amir Zeldes. 2018. All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proceedings of the Joint Workshop on Linguistic*

Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 167–177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *KONVENS 2016 Invited Talk*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Manuela Sanguinetti, Lauren Cassidy, Cristina Bosco, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2022. Treebanking user-generated content: A UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 57:493–544.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. A sequence-to-set network for nested named entity recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3936–3942. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2012. OntoNotes release 5.0. Technical report, Linguistic Data Consortium, Philadelphia.

Amir Zeldes. 2016. rstWeb - a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, San Diego, California. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes and Dan Simonson. 2016. Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 68–78, Berlin, Germany. Association for Computational Linguistics.

Shuo Zhang and Amir Zeldes. 2017. GitDOX: A linked version controlled online XML editor for manuscript transcription. In *Proceedings of FLAIRS 2017*, pages 619–623, Marco Island, FL.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

A Genre Descriptions

GENTLE comprises 8 genres, with each having 2 to 5 individual documents—cf. Table 2. They are as follows:

- **dictionary** – entries for a single English word from Wiktionary (`https://en.wiktionary.org`). GENTLE includes documents for the words *next*, *trust*, and *school*.

- **esports** – transcripts of a YouTube video clip containing esport commentary. GENTLE includes two documents: one featuring Fortnite, and the other featuring FIFA 20.

- **legal** – segments of legal text from the United States. Of the two documents, one is a portion of the Supreme Court opinion for Roe v. Wade (1973) from Wikisource (`https://en.wikisource.org`), and the other is a portion of a contract, extracted from the **C**ontract **U**nderstanding **A**tticus **D**ataset (CUAD) v1 from the The Atticus Project (Hendrycks et al., 2021).

- **medical** – snippets of a Subjective, Objective, Assessment and Plan (SOAP) note. A SOAP note is a common kind of text used by medical professionals to document a patient's medical visits and history. The notes are taken from MTSamples (`https://mtsamples.com`).

- **poetry** – poems taken from Wikisource (`https://en.wikisource.org/wiki/Portal:Poetry`). The poems come from 3 different authors and are of varying lengths.

- **proof** – mathematical proofs taken from ProofWiki (`https://proofwiki.org`).

- **syllabus** – syllabuses taken from course materials posted publicly on GitHub.

- **threat** – threat letters recorded in publicly available United States court proceedings. Accessed through casetext (`https://casetext.com/cases`; see also Abrams 2019 for some analysis of these texts).

B Full Label Residual Tables

The following tables give complete standardized Pearson residuals for label distributions in each GENTLE genre, along with comparisons to GUM as a whole and GUM news in particular. Tables 7–10 give numbers for UPOS, dependency, entity, and RST coarse labels respectively.

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
GUM	-0.5	-2.1	11.3	8.7	2.3	1.8	13.3	-15.1	-5.2	3.9	20.8	-25.1	-0.2	4.1	-14.6	6.6	-22.6
GUM_{news}	-1.3	5.9	-11.5	-6.2	-4.8	4.9	-11.5	5.7	4.6	-1.1	-21.7	41.4	-5.4	-3.3	-1.9	-3.3	-6.6
dictionary	7.9	-1.2	-5.9	-7.2	-0.4	-5.2	-2.9	4.4	-5.0	-0.6	-8.6	-4.9	20.4	-4.6	-2.1	-6.1	21.2
esports	-3.3	-1.1	5.9	2.1	-1.9	-0.6	0.4	-5.2	1.8	4.7	3.7	1.4	-4.1	-1.5	-1.3	4.2	-2.1
legal	0.3	0.0	-4.7	-5.5	3.0	3.2	-4.1	4.7	2.6	0.4	-9.5	3.9	0.8	0.0	4.9	-3.1	18.4
medical	7.6	0.0	-5.3	-0.4	0.0	-5.0	-4.0	11.6	5.4	-4.2	-2.6	-6.9	1.0	-3.3	-0.4	-4.8	8.9
poetry	-2.6	-1.4	5.0	-4.5	2.7	2.1	-2.8	-1.6	-5.3	-4.5	5.5	-7.1	5.5	0.9	-1.9	1.9	-2.2
proof	-1.0	0.4	0.5	0.3	-1.2	-4.7	-3.9	14.5	1.8	-5.8	-9.0	-8.1	0.9	3.7	56.3	-6.4	-2.3
syllabus	-1.4	-3.1	-6.6	-5.1	2.5	-6.4	-2.5	13.6	7.3	-4.5	-10.4	11.2	-3.4	-3.3	5.4	-5.3	54.5
threat	-2.5	-1.3	0.8	5.5	-0.5	-2.2	1.6	-2.6	-2.1	3.3	13.2	-7.4	-7.0	2.5	-1.7	4.5	-0.9

Table 7: Residuals for UPOS Labels by Genre.

	acl	advcl	advmod	amod	appos	aux	case	cc	ccomp	compound	conj	cop	csubj	dep	det	discourse	dislocated	expl
GUM	1.3	4.3	12.2	-1.9	-12.8	4.7	-4.3	2.3	-1.2	-16.1	-2.2	7.8	3.8	-20.9	1.4	9.9	-2.9	4.3
GUM_{news}	-0.6	-2.6	-11.7	3.5	7.5	-2.1	9.6	-5.1	4.9	20.3	-4.5	-7.1	-3.1	-2.5	5.3	-9.3	-1.4	-4.0
dictionary	-1.6	-4.4	-6.3	2.1	7.0	-5.2	-2.4	-0.5	-4.1	-2.2	9.7	-4.7	-1.7	5.6	-5.1	-3.1	-0.3	-2.4
esports	-1.5	0.9	4.7	-3.9	-2.3	2.3	-3.0	-2.4	-0.5	-0.4	-0.3	0.5	-1.3	-2.2	-0.7	2.1	9.2	-1.2
legal	4.2	-1.3	-4.4	1.4	3.6	-4.1	1.8	2.7	-2.7	3.4	1.8	-3.4	-1.4	17.2	3.3	-3.4	-0.2	-1.7
medical	-3.8	-4.3	-5.9	7.7	-1.3	0.4	0.2	0.1	-1.8	1.9	3.5	-0.9	-1.6	7.7	-4.7	-3.3	-0.2	-1.2
poetry	1.9	1.5	3.8	-2.8	-2.0	-4.3	-1.6	2.5	1.3	-5.7	-0.3	-1.8	-0.7	-3.6	2.2	-1.9	1.7	-0.7
proof	-1.2	-0.3	-0.1	-3.0	3.6	-3.2	1.6	0.5	1.6	-6.1	3.1	4.6	2.6	10.9	-5.2	-3.3	3.0	3.9
syllabus	-3.6	-2.9	-7.7	-0.5	19.8	-2.8	-3.5	2.7	-3.8	16.6	1.7	-4.5	-1.7	46.3	-6.5	-2.0	-0.3	-2.4
threat	2.5	2.4	3.0	-3.6	-2.8	5.9	-2.2	-0.7	2.2	-2.8	-0.0	1.0	0.4	-1.9	-2.1	3.3	1.7	0.5

(a) Part 1.

	fixed	flat	goeswith	iobj	list	mark	nmod	nsubj	nummod	obj	obl	orphan	parataxis	punct	reparandum	root	vocative	xcomp
GUM	0.0	-9.7	-2.7	0.6	1.8	5.0	-4.0	8.4	-7.2	4.6	-2.2	-0.1	-4.1	-0.1	5.5	-2.7	1.8	4.9
GUM_{news}	1.1	21.7	-0.5	-0.6	-1.6	-4.2	7.3	-4.2	4.2	-4.0	3.6	-0.8	-6.5	-5.4	-5.7	-6.8	-2.4	-4.8
dictionary	-0.3	-4.5	0.5	-1.7	-0.4	-2.2	-3.1	-9.8	-2.9	-5.3	-1.7	-0.0	18.4	20.4	-2.0	7.3	-0.3	-3.9
esports	1.5	-2.4	1.7	1.5	-0.3	3.9	-5.6	2.4	3.5	1.0	1.8	-0.4	16.7	-4.1	5.8	-1.5	-0.2	5.2
legal	-0.2	-1.6	0.6	-1.1	-0.3	-0.9	3.5	-6.6	0.6	-0.3	-1.5	-0.5	-2.0	0.8	-1.9	-2.4	-0.7	-2.6
medical	-1.2	-3.8	0.6	-1.5	-0.3	-3.8	3.4	-2.1	7.9	-2.6	0.5	2.2	-1.3	0.9	-1.8	6.7	-0.7	-1.8
poetry	-0.7	-2.8	0.6	1.6	-0.3	-2.2	-0.6	0.8	-2.2	0.1	0.5	-0.4	1.4	5.5	-0.1	0.1	1.7	-0.8
proof	-0.3	-4.5	0.6	-1.5	-0.3	0.0	2.7	1.4	-1.2	-1.2	2.1	0.1	-2.3	0.9	-1.8	1.3	-0.6	-1.1
syllabus	-0.3	1.1	0.5	-1.7	-0.4	-4.3	-3.1	-9.1	8.5	-1.9	-2.9	1.9	-0.5	-3.5	-1.7	16.4	-0.7	-3.9
threat	-1.5	-3.3	6.0	4.0	-0.3	2.6	-1.3	4.6	0.9	4.2	0.4	-0.4	1.1	-7.0	-1.6	-1.1	1.6	4.1

(b) Part 2.

Table 8: Residuals for Deprel Labels by Genre.

	abstract	animal	event	object	organization	person	place	plant	substance	time
GUM	-13.8	3.5	-3.1	6.4	-19.5	13.4	8.7	6.4	4.2	-2.3
GUM_{news}	-15.2	-7.7	7.2	0.5	30.6	-6.7	3.9	-2.2	-0.3	5.7
dictionary	17.1	5.0	-4.9	-5.9	6.6	-10.1	-6.1	-2.4	-3.6	0.4
esports	-9.3	-2.6	9.9	0.1	-0.5	7.8	-2.4	-2.0	-3.0	-0.7
legal	9.8	-2.7	1.9	-6.7	12.6	-8.0	-4.5	-2.1	-3.9	0.2
medical	4.4	-0.4	2.8	6.2	-6.1	-4.8	-7.8	-2.3	11.0	-0.2
poetry	-3.6	17.8	-4.3	1.5	-5.4	4.2	1.3	-0.8	-1.9	-1.4
proof	39.2	-3.0	-7.0	-7.7	-6.4	-16.5	-9.2	-2.4	-4.3	-6.7
syllabus	25.7	-3.4	-2.5	-8.6	-4.7	-12.1	-7.3	-2.7	-4.7	4.6
threat	-4.1	-2.2	-1.6	-1.0	-2.9	12.9	-2.9	-2.1	-3.7	-3.2

Table 9: Residuals for Entity Labels by Genre.

	adversative	attribution	causal	context	contingency	elaboration	evaluation	explanation	joint	mode	organization	purpose	restatement	same	topic
GUM	5.9	-3.7	2.6	0.6	1.3	-2.2	5.7	2.6	-8.9	1.9	-2.4	1.8	4.1	-0.8	7.7
GUM_{news}	-3.6	12.1	2.1	3.6	-1.8	5.3	-4.4	-6.0	-2.7	-2.1	-4.6	2.2	-3.5	1.4	-5.2
dictionary	-3.6	-5.4	-4.3	1.0	-2.3	2.6	-4.0	0.6	3.2	-1.8	5.7	-2.5	0.1	3.8	-2.6
esports	0.3	-1.2	-0.2	1.5	-1.6	-3.2	6.8	-1.2	1.3	-0.2	-0.8	0.9	0.3	-0.3	-1.4
legal	-2.5	-3.6	-2.3	-3.7	0.3	3.1	-3.0	0.9	2.3	-1.4	2.1	1.5	0.2	2.9	-1.9
medical	-1.6	-2.2	-2.4	-3.6	-1.5	-4.2	-3.2	-3.5	12.9	-1.4	9.1	-2.5	-1.9	-3.3	-2.0
poetry	2.4	1.7	2.7	0.3	-2.2	-0.3	-0.1	-1.9	-3.1	5.5	-3.0	-2.3	2.7	2.5	0.0
proof	-4.2	1.6	-2.1	2.4	4.2	-1.4	-3.0	8.5	-0.0	0.4	2.7	-2.5	-2.3	-3.0	-2.0
syllabus	-3.3	-5.1	-4.3	-5.5	-0.5	-0.7	-4.1	-4.0	18.2	-1.8	4.3	-2.7	-3.5	-2.0	-2.6
threat	1.2	1.1	0.5	-2.2	4.1	-0.8	2.4	6.4	-1.3	-0.5	-2.9	-0.3	-1.3	-1.5	-2.0

Table 10: Residuals for RST Relation Classes by Genre.