Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs

Natalie Hervieux[1*] **Peiran Yao**[1*] **Susan Brown**[2] **Denilson Barbosa**[1]
[1]University of Alberta [2]University of Guelph
{nhervieu,denilson}@ualberta.ca

Abstract

The digital humanities (DH) community fundamentally embraces the use of computerized tools for the study and creation of knowledge related to language, history, culture, and human values, in which natural language plays a prominent role. Many successful DH tools rely heavily on Natural Language Processing methods, and several efforts exist within the DH community to promote the use of newer and better tools. Nevertheless, most NLP research is driven by web corpora that are noticeably different from texts commonly found in DH artifacts, which tend to use richer language and refer to rarer entities. Thus, the near-human performance achieved by state-of-the-art NLP tools on web texts might not be achievable on DH texts. We introduce a dataset[1] carefully created by computer scientists and digital humanists intended to serve as a reference point for the development and evaluation of NLP tools. The dataset is a subset of a born-digital textbase resulting from a prominent and ongoing experiment in digital literary history, containing thousands of multi-sentence excerpts that are suited for information extraction tasks. We fully describe the dataset and show that its language is demonstrably different than the corpora normally used in training language resources in the NLP community.

1 Introduction

The digital humanities (DH) research community makes up a large user base for natural language processing (NLP) tools and algorithms (McGillivray et al., 2020; Biemann et al., 2014). Digital humanists have long been using cultural heritage data for meaningful NLP work, where NLP in DH includes everything from linguistic analysis of change over time within large linguistic corpora (Schlechtweg et al., 2020) to narratology (Piper et al., 2021) to literary history (Underwood et al., 2018) and stylometry (Stamatatos, 2009).

However, there are risks associated with LLMs that are particularly relevant to DH. Unlike the average web document, texts in the humanities tend to use rich and complex writing styles, historical language, and references to under-represented long-tail entities (Olieman et al., 2017; Nurmikko-Fuller, 2023). LLMs have known problems with bias towards the contemporary and popularity bias (Dai et al., 2024). Chen et al. (2024) warn of a "Spiral of Silence" where over time, by iteratively training on LLM-generated content, LLM-based retrieval systems deprioritize accurate human-generated content and lose diversity in the information they return. If future NLP is dominated by LLMs that ignore the outliers that are so important to humanities scholarship (D'Ignazio, 2021; Jockers, 2013), this will negatively impact humanities research, our sense of history, and the public. As Brown and Simpson (2013) assert, "marginality and uniqueness are what humanities scholars often seek to discover and analyse". We need curated datasets for evaluating and fine-tuning LLMs with the priorities and expertise of humanists at their core.

For LLMs to effectively and responsibly leverage this data and become reliable for DH needs, researchers developing these models and the tools that use the models, need to collaborate with data experts. As McGillivray et al. (2020) point out, there is a need for cross-fertilization of ideas and more communication across the NLP and DH communities. LINCS (Brown et al., 2023) is an example of computer science (CS) and humanities practitioners working together to extract knowledge from DH texts in the form of linked data connected to web pages to create machine-readable data that could ultimately enhance LLMs. However, the inability of current systems to handle the ontological nuances of the source data plus the absence of entities from popular knowledge bases (KBs) like

[*]Contributed equally to this work.
[1]https://doi.org/10.5683/SP3/RCVANO

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 85–104
November 16, 2024.

Wikipedia and Wikidata (Vrandecic and Krötzsch, 2014) necessitates manual entity linking, ontology mapping, and data validation. There is an opportunity here for NLP developers to better support such projects with systems optimized for cultural heritage data.

To contribute to these efforts, **we create an NLP dataset through a collaborative effort between computer scientists and humanists**. Our dataset, *Orlando (Release)* (Hervieux et al., 2024), consists of 12,627 unique text chunks with over 40,000 entity mentions across four entity types that are manually linked to external entity URIs and annotated with 79 unique inter- and cross-sentence relations. The source texts are biographies of historical writers from a large and representative born-digital humanities corpus created by the Orlando Project (Brown et al., 2022) (Appendix A). These source biographies are originally expressed as XML documents, written and thoughtfully hand-annotated in English by DH scholars, using language demonstrably richer than that found in typical LLM training corpora. We extract our dataset from the source while ensuring a high rate of long-tail entities, and preserving the ontological nuances of the source texts' entity and relationship annotations, which we augment with manually-confirmed entity URIs (§3). This makes the dataset particularly well-suited for information extraction tasks such as entity linking (EL) and relation extraction (RE), as elaborated in §5.

Orlando's text complexity (examples in Appendix A.1) makes it an interesting subject of study for what machine-aided tools can process. We conduct a series of linguistic analyses (§4) to show that, compared to other genres of text such as news, encyclopedia, or web pages, the Orlando data is more complex in terms of both lexical and syntactic aspects. In light of that, we test whether the Orlando data is out-of-distribution for state-of-the-art large language models such as Llama 2 (Touvron et al., 2023), using metrics based on Mahalanobis distance (Ren et al., 2023) and kernel density estimation (Kirchenbauer et al., 2024). The test highlights that the Orlando data is evidently further from the distributions of the training corpora of LLMs than baseline corpora. This suggests that LLMs, when used out-of-the-box, may suffer from poorer generalization, lower accuracy and higher aleatoric uncertainty (Baan et al., 2023) when processing complex DH text like that from Orlando.

We hope that this data will encourage NLP tool developers to embrace the challenges posed by DH texts and seek collaboration with the data experts, leading to research, data, tools and systems that would be valuable across disciplines.

2 Related Work

Our dataset is a unique addition to the important yet disproportionally scarce collection of information extraction datasets derived from humanities texts and created collaboratively by NLP and DH researchers. There are countless datasets created to benchmark information extraction models (Nasar et al., 2022) and many works that perform such benchmark evaluations (Wang et al., 2022; Chang et al., 2024), but they typically lack the humanities perspective. There are exceptions, with examples including but not limited to Menini et al. (2022)'s information extraction benchmark relevant to cultural historians interested in textual descriptions of smells in historical documents; Delmonte and Busetto (2023)'s investigation of BERT's limitations when applied to linguistically complex italian poetry; Pedinotti et al. (2021)'s diagnostic dataset and evaluation of transformer-based language models on generalized event knowledge; and Bamman et al. (2020)'s challenging coreference resolution dataset for literary texts. These works focus on other genres of text than that of Orlando, and our domain allows us to provide hand-curated cross-database annotations for entity mentions, which is crucial for the evaluation of EL. Compared to works that evaluate BERT and other specialized models, we focus on the suitability of LLMs for humanities-related information extraction tasks. This is critical as LLMs are becoming the status quo for many NLP tasks (Chang et al., 2024) and LLMs are often used with a different paradigm: zero-shot prompting rather than fitting to the target domain.

There are many other valuable datasets coming out of the humanities[2,3,4]. Major differences between these datasets and ours are that most of these projects release their entire research corpora as raw text with humanities research as the target task rather than information extraction or LLM benchmarking or fine-tuning. Our approach was to look at a prominent humanities dataset with challenging language, consult with DH scholars to understand

[2] https://rutgersdh.github.io/dh-sources/
[3] https://humanitiesdata.com/resources
[4] https://melaniewalsh.github.io/Intro-Cultural-Analytics

what level of information was important to keep, and then selectively sample it with the intent of keeping difficult chunks that contain many diverse entities and relations.

Our text analysis of Orlando draws on work in evaluating text readability (Crossley et al., 2011; Lu, 2010), but our work deviates as we apply readability measures to compare corpora used in NLP models. We use the popular Flesh-Kincaid grade level (Kincaid et al., 1975) which suits our chunk-level data compared to other metrics like Coh-Metrix (Graesser et al., 2004) which requires paragraph statistics and discourse coherence. See Lu's (2014) work for a corpus linguist's review of computational corpus analysis. Our work interrogates whether a corpus is out-of-distribution of an LLM's training corpus, picking the best-performing indicators in recent discussions (Ren et al., 2023; Yauney et al., 2023; Kirchenbauer et al., 2024).

3 Creating the Orlando Dataset

Derived from the original Orlando XML documents, we release a simplified and easily machine processable JSON dataset, Orlando (Release). Through this collaboration with the data experts, we simplify the complexly nested embedded annotations into an easy-to-use benchmark, without abstracting the nuance of the original entity and relation types.

According to the license, we can release 10% of the Orlando textbase. Instead of uniformly sampling from all sentences or entire documents, we release text chunks of up to 4 sentences each that capture valuable cross-sentence relationships and helpful context for coreference resolution and EL tasks. We filter out text chunks containing too few entities or relations to ensure a high density of useful text. When our extractions come from overlapping chunks, we merge smaller ones into larger ones. We select the included text chunks randomly but with constraints to keep the original frequency distribution for relations and to prioritize the inclusion of person mentions with external entity links. Our sampling process does not alter the distribution of data as it is uniform sampling in a stratified fashion that preserves the long tail distributions of relation and entity types.

Orlando (Release) has 12,627 unique text chunks with over 40 thousand entity mentions across four entity types, with the majority being person mentions. Table 1 lists entity mention counts by type and presence of external entity links, and breaks down person mentions into the primary subjects of the biographies and other mentioned persons.

	Mentions	Entities
person		
bio subjects	14,168	1,389
bio subjects with links	14,122	1,379
others	10,627	6,257
others with links	6,951	3,145
organization	2,910	1,466
place	11,638	4,785
creative work	1,127	928

Table 1: Mention and unique entity count for each entity type in Orlando (Release). Place and creative work types were not de-duplicated so entity count is the number of unique mentions. "with links" rows are subsets of the row directly above.

Compared to typical RE benchmarks with few broad relations, our dataset contains 79 unique relations, 30 of which are present in at least two contextual categories. The full list of relations and categories with frequency statistics are in the Appendix B Tables 8, 9, and 10.

3.1 Source Textbase

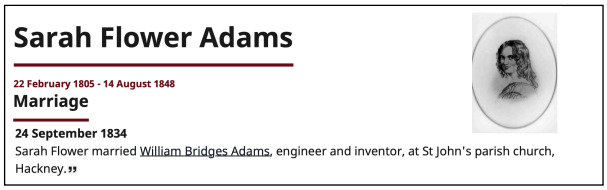
The original Orlando documents are densely annotated XML biocritical profiles of authors (*biographies*). Tags are applied on the word level to identify and add context to entities and concepts, and on the sentence or paragraph level for contextual themes and relations. Figure 1 presents an example. The annotations signal what is most relevant to the domain researchers, which means that not all possible entities and relations are tagged. The data is unique in that the included annotations are extremely detailed, as we discussed in noting the wide range of relations.

Each biography focuses on one person, who we refer to as the biography subject. The biographies are sectioned in two: (1) their birth, death, and the people, places, and activities in between; (2) their writing and its reception. We sample only from the first to prioritize capturing relations between persons without the added complexity of written and often fictional works.

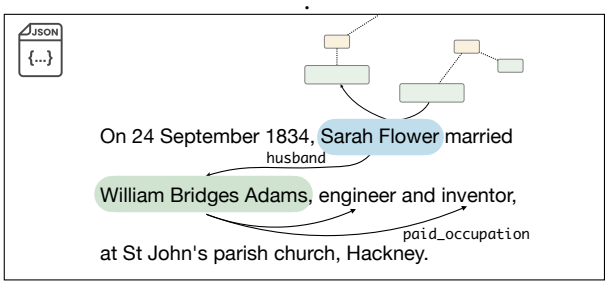
Entity Tags There are four entity types explicitly tagged in the XML documents that we include in our dataset: *person*, *place*, *organization*, and *creative work*. Pronouns are not flagged by annotators

```
<HEADING>Marriage</HEADING>
<FAMILY>
  <MEMBER RELATION="HUSBAND">
    <MARRIAGE>
      <CHRONSTRUCT RELEVANCE="COMPREHENSIVE">
        <DATE VALUE="1834-09-24">24 September 1834</DATE>
          <CHRONPROSE>
            <NAME STANDARD="Adams, Sarah Flower"
                  REF="...d681ef">
              Sarah Flower
            </NAME> married
            <NAME STANDARD="Adams, William Bridges"
                  REF="...9f83e2">
              William Bridges Adams
            </NAME>,
            <JOB>engineer</JOB> and
            <JOB>inventor</JOB>, at
            <PLACE>
              <PLACENAME REG="St John's Church, Hackney">
                St John's parish church</PLACENAME>,
              <SETTLEMENT CURRENT="London">Hackney</SETTLEMENT>
              <REGION REG="Middlesex"/>
              <GEOG REG="England"/>
            </PLACE>.
          </CHRONPROSE>
        </CHRONSTRUCT>
      </MARRIAGE>
    </MEMBER>
</FAMILY>
```

(a) The original Orlando documents are XML in which text is embedded within deeply-nested relation and entity tags.



(b) Author profile corresponding to the XML in Figure 1a.



(c) We release JSON documents with the same information as the XML but with text extracted and cleaned and lists of relations and entities enhanced with external identifiers.

Figure 1: Representation of an Orlando text chunk displayed as its source XML document (a), its published form on the web (b), and its extracted form in our released dataset (c).

and thus do not appear as mentions in our data. However, there are cases where a person's relationship to another is used as the mention. For example, "Elizabeth Singer Rowe returned to Frome to live with her father" contains the mention "father". We include such mentions because a human annotator could confirm a match using the context, so a sophisticated EL method may also be able to.

Relation Tags The XML relation tags indicate how something in the text relates back to the biography subject, making them the subject of all extracted relations. There is, therefore, no specific text span to connect them to a given relation. The only exception is that the occupations of family members are explicitly tagged.

Rather than tagging specific verbs to represent explicit relations, the annotators tag multi-sentence and sometimes multi-paragraph chunks with specific categorical terms[5]. For each category, there are certain nested tags that we use to extract relations. For example, within a <FAMILY> tag, there can be a nested <MARRIAGE> element, within which the first <NAME> element represents the biography subject's spouse, and within <DEATH>, we may find <DATE>, <CAUSE>, and <PLACE> with details of the biography subject's death.

[5] https://orlando.cambridge.org/index.php/about/tag-diagrams

3.2 Extracted Dataset

Pre-processing We apply an automated text-cleaning step before extracting our dataset from the XML, correcting typographic whitespace errors and integrating dates at the starts of sentences rather than as headings. Originally, the biography subjects are mentioned with project-specific acronyms, which we replace with full names as defined by the annotators. As such, a subject is always mentioned with the same text, except for female subjects when called their birth name early on and their married name later.

Finding Entity Links 20 annotators with backgrounds in humanities and CS[6] manually searched, using OpenRefine (Delpeuch et al., 2024), for external identifiers for a subset of the over 27,000 unique person entities. To get a broad sample, the review began with the first 8,500 persons by alphabetical order, then 3,240 remaining persons with the highest mention count across all biographies. All biography subjects had been previously searched for by earlier annotators who also found matches for many places in GeoNames (Unxos, 2013).

We instructed annotators to choose one match per entity from VIAF (Tillett, 2002), Wikidata, or Getty ULAN (Harpring, 2010). The Orlando Project leads deemed those sources useful for creat-

[6] The annotators were undergraduate students enrolled in humanities and CS programs and were paid employees of the authors' universities.

ing meaningful linked open data. Using any available context from the biographies or the web, annotators confirmed matches when they were "definite" or "reasonably certain" on our four-point scale. These scores required multiple pieces of evidence, such as matching birth and death dates, titles of written works, or family members. When annotators could not confirm a match, either because of inadequate evidence or absence of a viable candidate, our data specifies "unable to confirm match." If an entity was not reviewed by our annotators, we mark it as "match not searched for." Using the confirmed matches, we query Wikidata's SPARQL endpoint to get equivalent URIs across the three sources and Wikipedia. All found *links* are included in our dataset to facilitate benchmarking systems that use different KBs.

4 Corpus Comparison

We compare Orlando with corpora of varying genres to determine its complexity for human readers and automated processing.

4.1 Baseline Corpora

We select baselines by two criteria. They spread across diverse genres, including news, encyclopedia, and webpages. They also represent the typical corpora used in training LLMs to provide a more accurate projection of the difficulty of Orlando for LLMs. Each corpus is pre-processed using the same pipeline as Orlando (detailed in Appendix C.1). The corpora are:

C4 Common Crawl[7] is a large corpus of webpages, reflecting the proportions of different textual content available on the Internet. We use the derived C4 dataset (Raffel et al., 2020a), a cleaned version of Common Crawl that only contains English webpages, as it is the backbone training corpus for many LLMs (Raffel et al., 2020b; Chalkidis et al., 2022; Groeneveld et al., 2024).

CC-News We use the subset of CC-News (Nagel, 2016) prepared by Liu et al. (2019) using *news-please* (Hamborg et al., 2017), which is a dataset of 708,241 English-only news articles extracted from Common Crawl. It is part of the mixture of training corpora of smaller scale language models such as RoBERTa (Liu et al., 2019).

[7] https://commoncrawl.org

Corpus	FKGL	Avg. Entities
C4	9.56	1.13
CC-News	9.66	1.88
Wikipedia	11.75	2.84
Simple Wiki	8.93	2.16
Orlando (Full)	11.47	2.40
Orlando (Release)	11.90	3.15

Table 2: Flesch-Kincaid Grade Level (FKGL) and average number of entities per sentence of the corpora.

Wikipedia The English Wikipedia is a large encyclopedia that is also widely used as a training corpus for a full spectrum of language models as summarized by Alshahrani et al. (2023).

Simple Wiki As a reference point for text complexity, we include Simple English Wikipedia. It is a version of Wikipedia that is written in simple English and is intended for people with different language proficiency levels.

Orlando (Full) As a baseline, our comparison includes the full text of all biographies in the Orlando Project (Brown et al., 2022). The dataset we release, designated as **Orlando (Release)**, is a subset of the full Orlando dataset that only contains high-quality chunks of text satisfying the criteria described in §3.

4.2 Lexical Complexity

Metrics We count the number of entities in each sentence and report the Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) in Table 2. FKGL maps the number of syllables and words in a sentence to the number of years of education required to understand the sentence, and is widely used in the automatic evaluation of text complexity (Alva-Manchego et al., 2019).

Discussion Overall, Orlando is among the most complex corpora in terms of lexical complexity which could pose difficulties for human readers. The similarity between Orlando and Wikipedia is expected, as they share a similar genre with biographical text comprising a large part of Wikipedia. They contain more named entities per sentence than the other corpora. In particular, the distribution of the number of entities in our released subset skews towards the right, with the highest mean.

From an information extraction perspective, the high number of entities per sentence makes Orlando harder to process as it requires more EL,

coreference resolution, and RE operations. The released subset is selected with a preference to contain sentences with more entities and relations, which makes it more suitable for EL and RE benchmarking and leads to a more challenging dataset.

4.3 Syntactic Complexity

Metrics We use the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010) to analyze the syntactic complexity of the corpora. This widely adopted tool enumerates a list of patterns in a parse tree and produces 13 variables associated with five aspects of syntactic complexity: length, subordination, coordination, overall complexity, and phrasal sophistication. We plot the scores of the corpora concerning each of the five aspects in Figure 2.

Discussion L2SCA shows that Orlando has higher syntactical complexity than the other corpora. It ranks high in all five aspects of syntactic complexity, with the highest scores in length of production units and amounts of coordination. In comparison, CC-News has high number of subordinations, but fewer coordination and shorter production units. Wikipedia has more subordinations than CC-News and the highest ratio of complex nominals, but it has fewer coordination and shorter production units.

4.4 In-distribution Assessment

Metrics There is mounting theoretical (Saunshi et al., 2021) and empirical (Razeghi et al., 2022; Kandpal et al., 2023; Ren et al., 2023; Kirchenbauer et al., 2024) evidence that suggests a positive correlation between the similarity of the distributions of training and test data and LLM's performance. Therefore, assessing whether a test dataset (Orlando in our case) is in-distribution, i.e. it follows the same distribution of a model's training distribution, could be indicative of the model's relative performance on the Orlando dataset.

We adapt two metrics to measure whether Orlando is in-distribution: Mahalanobis distance (MD; Ren et al., 2023) and kernel density estimation (KDE; Kirchenbauer et al. 2024) with respect to the training data distribution. The metrics are shown to be correlated with model's performance on translation and language understanding respectively. Both methods represent training and test samples in the embedding space. Ren et al. (2023) fits the training data to a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ and computes the squared Mahalanobis distance $MD(x) = (x-\mu)^T\Sigma^{-1}(x-\mu)$ for each test sample. Kirchenbauer et al. (2024) takes a non-parametric approach and estimates the probability density of each test sample directly from the training data using the approximate KDE algorithm by Karppa et al. (2022).

The two metrics both require access to the training data of a model, while MD also relies on the internal activations of LLMs. The former is generally unavailable except for works from the LLM open-science community such as (Soldaini et al., 2024) and (Groeneveld et al., 2024). The latter is also unavailable for blackbox LLMs such as GPT-4. For MD, we analyze two open-weight LMs: decoder-only `Llama-2-7B` (Touvron et al., 2023) and encoder-decoder `BART-large` (Lewis et al., 2020), and assume C4 to be a good approximation of the training data based on the observations in §4.1. For KDE, we use Soldaini et al.'s (2024) open-science replica to approximate frontier LLMs' training data.

Discussion As shown in Figure 3, both MD and KDE show that Orlando as test data has lower density in LLM's training data distribution, indicating that Orlando contains more long-tail information (to be discussed in §4.5) and is more likely to be out of the distribution, compared to general webpages, news or Wikipedia articles. While existing research does not establish a clear density threshold for ensuring the acceptable performance of LLMs, the findings indicate a need for extra caution, as the use of LLMs with DH data may lead to relatively degraded performance. As a future direction, we suggest more directly benchmarking LLMs on DH datasets such as Orlando.

4.5 A High Percentage of Long-tail Entities

Thanks to our URI attribution, we find the percentage of Orlando (Full) person entities in common KBs: Wikipedia for notable people, Wikidata as a larger and more diverse KB, VIAF for people with publications which are relevant to Orlando, and Getty ULAN as an example relevant to many humanities texts but less so Orlando.

Table 3 presents the results for 1,434 subjects and 8,510 randomly sampled other people. Unsurprisingly, over 90% of subjects notable enough to have biographies written about them are found in each of the three relevant KBs. However, 50.8%, 41.1%, and 37.5% of the other people could not be found in Wikipedia, Wikidata, and VIAF, respec-

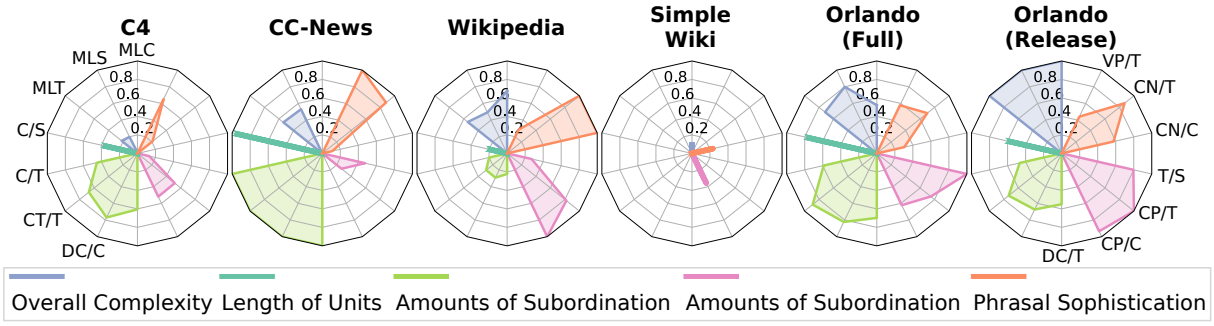
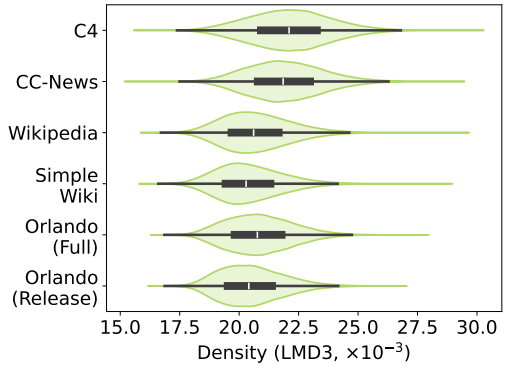


Figure 2: L2SCA values of the six corpora, with each corpus plotted as an individual circle. Within a circle, each polygon represents an aspect of L2SCA and each vertex represents a variable. Starting from 12 o'clock and iterating counterclockwise, the respective aspects of the polygons are listed in the legend. The area of the polygon is proportional to the scores of the aspect. For illustration, the values are normalized by min-max scaling. Definitions of the abbreviated variable names and raw values of the 14 variables of L2SCA (Lu, 2010) in Tables 11, 12 and 13.

	C4	CC-News	Wikipedia	Simple Wiki	Orlando (Full)	Orlando (Release)
Llama-2 -7B	0.0	10.2	23.9	30.6	187.2	261.1
BART -large	0.0	9.6	15.7	23.3	102.1	134.6

(a) Mean squared Mahalanobis distance.



(b) Kernel density estimation.

Figure 3: Both MD and KDE show that Orlando is relatively more out-of-distribution. Higher chance of samples being out-of-distribution results in higher MD and lower KDE.

tively by our human annotators – either because the entity was not present or there was insufficient evidence to make a match. These rates highlight that a large percentage of Orlando entities are not considered notable and demonstrate Orlando's high concentration of long-tail entities (Kandpal et al., 2023).

Kandpal et al. (2023) explore the relationship between question answering performance and the number of documents about an entity in the training data, and report reduced performance in connection with long-tail entities. This raises questions about how LLMs and associated tools will perform on entity-based tasks with data such as Orlando. There is an opportunity here for LLMs to harness more humanities data to work better for long-tail entities – ultimately reducing historical biases and uplifting historically silenced and overlooked individuals. It also highlights the importance of datasets such as ours so that systems can be evaluated on a mix of popular and long-tail entities.

	Bio Subjects	Others
Wikipedia	93.6%	49.2%
Wikidata	98.7%	58.9%
VIAF	94.3%	62.5%
ULAN	13.2%	8.0%

Table 3: Percentage of unique person entities reviewed by annotators that have matches in each KB. This is a sample of 1434 biography subjects and 8510 others mentioned in Orlando (Full).

5 Exploring Dataset Use Cases

The unique linguistic features of Orlando texts have made them subjects of study in applications like text simplification (Yao et al., 2024). Our dataset enriches the texts with annotations focused on entity and relation mentions, making it well suited for information extraction tasks. We demonstrate the data's usefulness through off-the-shelf EL and RE systems simple to use without customization.

Entity Linking We use the zero-shot EL system BLINK (Wu et al., 2020, details in Appendix D.1), which uses transfer learning and is potentially useful when applied to the humanities because it should not require training data from the target do-

	Bio Subjects	Others
By Entity	0.89	0.80
By Mention	0.92	0.81

Table 4: BLINK entity linking accuracy on Orlando (Release) using BLINK's pre-trained Wikipedia model.

main. We link the 13,727 mentions of the 1,307 unique biography subjects and the 5,920 mentions of the unique 2,528 other persons that have confirmed Wikipedia links in Orlando (Release).

Many EL systems consist of an end-to-end pipeline for both recognition (finding mentions to entities) and linking (matching each mention to a database entry). Orlando (Release) enables the evaluation of both steps, but we limit this evaluation to the linking step because BLINK uses a third-party named entity recognition (NER) system. As such, and as is standard in this setting where the system is not able to abstain from making a prediction, we report only accuracy[8] (Wu et al., 2020; Botha et al., 2020; Hoffart et al., 2011).

Table 4 presents the accuracy broken down by mention and unique entity for each person type. The two rows "By Mention" and "By Entity" refer to two common ways to aggregate results in EL literature. "By Mention" accuracy is micro-averaged as in Hoffart et al. (2011) or the number of correct matches divided by the number of mentions. "By Entity" is the macro-average, calculated as the number of correct matches divided by the number of mentions of entity e, for each entity e in our dataset and then taking the average.

BLINK performs similarly on Orlando subjects to what Wu et al. (2020) report on TACKBP-2010 (0.92 here compared to their best accuracy of 0.945). However, we see a 0.09 to 0.11 point decrease between the notable subjects set and that of the other people, which contains more long-tail entities. It is also important to note that, following Wu et al. (2020)'s problem setup and because of BLINK's inability to make NIL prediction, we report accuracy only of the entities for which we have confirmed Wikipedia links. This results in artificially inflated scores that are not reflective of the reality of EL on humanities texts. These issues highlight the potential our data has as a challenging entity linking benchmark.

[8]Accuracy and precision are equal in this task setting.

Relation Extraction We use the end-to-end RE system PURE (Zhong and Chen, 2021) on a random sample of 50 text chunks from Orlando (Release), pre-processed as described in Appendix D.2. PURE uses a small set of generic predicates based on those used in the ACE05 dataset (Walker et al., 2006), where a predicate is the connecting term in the subject-predicate-object representation of an extracted relation. On this sample, PURE's results include six unique predicates, while Orlando (Release) includes 34. Table 5 shows our mapping between Orlando and PURE predicates for the relations that were correctly present in both the Orlando annotations and in PURE's results for this sample.

PURE	Orlando
General-Affiliation	relocation
Person-Social	brother, husband, interpersonal_relationship, instructor
Physical	habitation, relocation, travel, visit

Table 5: The mapping between PURE predicates and Orlando predicates on the relations that both PURE and Orlando correctly identify on a sample of 50 random Orlando (Release) text chunks. This is only 8 of 34 unique Orlando predicates from this sample that PURE found equivalents to.

Even in this small sample of overlapping predicates, we see PURE abstracting away the valuable specificity in Orlando's thoughtfully created predicates. Of course there are systems with predicate sets ranging in size and specificity, but we use PURE as a demonstration that our data can be used to evaluate and improve systems across that range.

We manually verify each relation in PURE's results. Of the 115 relations that PURE extracts, 83% are correct and 65% are both correct and not found in our Orlando annotations. However, the predicates that PURE uses are so high-level that it is challenging to derive meaning from many of the new extractions. PURE only finds 10% of the 174 annotated relations in this sample. This indicates that the detail contained in Orlando poses a significant challenge for such RE systems. Table 6 provides an example, showing the relations that PURE finds for the text in Appendix D.3 that

subject	predicate	object
her	Person-Social	parents
student	Organization-Affiliation	school
her	Organization-Affiliation	school
school	General-Affiliation	Canada
Annie Louisa Walker	social_class	professional but not wealthy rank among the middle classes
Annie Louisa Walker	nationality	English
Annie Louisa Walker	race_colour	white
Annie Louisa Walker	religion	Christian
Annie Louisa Walker	religion	Evangelicals
Annie Louisa Walker	gendered_political_activity	Temperance movement
Annie Louisa Walker	political_involvement	Temperance movement

Table 6: An example of PURE (top) and Orlando (Release) (bottom) relations on the same text sample. PURE abstracts away Orlando's valuable detailed predicates.

subject	predicate	object
Philip Larkin	school	St John's College, Oxford
Philip Larkin	subject_studied	English language and literature
Philip Larkin	degree	Honours BA
Philip Larkin	education_companion	Bruce Montgomery
Philip Larkin	education_companion	Kingsley Amis
Philip Larkin	contested_behaviour	Amis and Larkin constituted themselves a two-man parody factory mocking every aspect of university life: the syllabus, the dons, and the aspiring writers such as John Heath-Stubbs.

Table 7: An example of Orlando (Release) relation annotations on text where PURE was not able to identify any relations.

are technically correct, but that lack specificity – even if we were to incorporate coreference resolution into the results. Table 7 shows the detailed and varied relation annotations included in Orlando (Release) for the text in Appendix D.4 on which PURE returns no results.

6 Conclusion

We argue that the impressive results reported by fast-paced NLP research might not reach tools in the DH community due to inherent differences in the kinds of texts they use. In particular, we note that LLMs have been shown to underperform with out-of-distribution inputs compared to experiments where test data comes from the same distribution as the training data (which is the norm in NLP research). While there is currently no machinery to predict the gap in observed performance for a given dataset, we report statistics derived using state-of-the-art methods that indicate noticeable differences between a corpus derived from a prominent born-digital DH textbase and corpora commonly used as training data in NLP research. We contribute this collaboratively developed dataset and argue for its potential to help close the gap between DH scholars and NLP system developers by serving as a benchmark for existing (and future) tools, as well as a resource for tool development.

Limitations

While we provide extensive statistical analysis using state-of-the-art methods, we consider only one (albeit prominent) DH dataset. Many avenues for future work exist. First, a similar analysis with a larger sample of prominent texts from the DH community, covering a range of genres, is needed. From a tool development point of view, an immediate use of our dataset would be fine-tuning

existing large language models to improve their ability to handle similar texts. We hope that our data will also be used in the evaluation and development of NER, EL, and RE tools that are better equipped to handle rich and complex texts with mentions of rarer entities compared to news and other kinds of texts found on the web. More importantly, we see tremendous potential in using other extant resources from the DH community in the development and evaluation of NLP tools. Such an approach can only lead to more robust tools.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Foundation for Innovation.

We thank LINCS[9], CWRC[10], and the Orlando Project[11] for funding students to complete the entity link attribution part of this work, and we are grateful to those students for their contributions.

References

Saied Alshahrani, Norah Alshahrani, and Jeanna Matthews. 2023. DEPTH+: An enhanced depth metric for Wikipedia corpora quality. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 175–189, Toronto, Canada. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *CoRR*, abs/2307.15703.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in english literature. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 44–54. European Language Resources Association.

Chris Biemann, Gregory R Crane, Christiane D Fellbaum, and Alexander Mehler. 2014. Computational humanities-bridging the gap between computer science and digital humanities (dagstuhl seminar 14301). In *Dagstuhl reports*, volume 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA 95472, USA.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Susan Brown, Patricia Clements, and Isobel Grundy. 2022. Orlando: Women's writing in the british isles from the beginnings to the present.

Susan Brown, Kim Martin, and Asen Ivanov. 2023. "Linking Out: The Long Now of DH Infrastructures.". In Paul Barrett and Sarah Roger, editors, *Future Horizons: Canadian Digital Humanities*, chapter 18. University of Ottawa Press, Ottawa, Ontario K1P 6B9, Canada.

Susan Brown and John Simpson. 2013. The curious identity of michael field and its implications for humanities research with the semantic web. In *2013 IEEE International Conference on Big Data (IEEE BigData 2013), 6-9 October 2013, Santa Clara, CA, USA*, pages 77–85. IEEE Computer Society.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022. An exploration of hierarchical attention transformers for efficient long document classification. *CoRR*, abs/2210.05529.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*

Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024. Spiral of silences: How is large language model killing information retrieval? - A case study on open domain question answering. *CoRR*, abs/2404.10496.

Scott A Crossley, David B Allen, and Danielle S McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1):84–101.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6437–6447. ACM.

[9] https://lincsproject.ca/
[10] https://cwrc.ca/
[11] https://orlando.cambridge.org

Rodolfo Delmonte and Nicolò Busetto. 2023. Stress test for BERT and deep models: Predicting words from italian poetry. *CoRR*, abs/2302.09303.

Antonin Delpeuch, Tom Morris, David Huynh, Weblate (bot), Stefano Mazzocchi, Jacky, Thad Guidry, elebitzero, Owen Stephens, Isao Matsunami, Iain Sproat, Albin Larsson, Silvério Santos, allanaaa, kushthedude, Sandra Fauconnier, Ekta Mishra, Martin Magdinier, Antoine Beaubien, Lu Liu, Joanne Ong, Fabio Tacchelli, Florian Giroud, Allan Nordhøy, Luca Martinelli [Sannita], Elroy Kanye, Mathieu Saby, and Lisa Chandra. 2024. Openrefine/openrefine: 3.8.2.

Catherine D'Ignazio. 2021. Outlier. In Nanna Bonde Thylstrup, Daniela Agostinho, Annie Ring, Catherine D'Ignazio, and Kristin Veel, editors, *Uncertain archives: Critical keywords for big data*, chapter 40. MIT Press.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *CoRR*, abs/2402.00838.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please - A generic news crawler and extractor. In Maria Gäde, Violeta Trkulja, and Vivien Petras, editors, *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science, ISI 2017*, volume 70 of *Schriften zur Informationswissenschaft*, pages 218–223. Verlag Werner Hülsbusch, Berlin, Germany.

Patricia Harpring. 2010. Development of the getty vocabularies: Aat, tgn, ulan, and cona. *Art Documentation: Journal of the Art Libraries Society of North America*, 29(1):67–72.

Natalie Hervieux, Peiran Yao, Susan Brown, and Denilson Barbosa. 2024. Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Matti Karppa, Martin Aumüller, and Rasmus Pagh. 2022. DEANN: speeding up kernel-density estimation using approximate nearest neighbor search. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 3108–3137. PMLR.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

John Kirchenbauer, Garrett Honke, Gowthami Somepalli, Jonas Geiping, Katherine Lee, Daphne Ippolito, Tom Goldstein, and David Andre. 2024. LMD3: Language model data density dependence. In *First Conference on Language Modeling*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2014. *Computational Methods for Corpus Annotation and Analysis*. Springer Dordrecht, DORDRECHT, Netherlands.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz Fabo. 2020. Digital humanities and natural language processing:"je t'aime... moi non plus". *Digital Humanities Quarterly*, 14(2).

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke van Erp, Inger Leemans, Pasquale Lisena, Raphaël Troncy, William Tullett, Ali Hürriyetoglu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic, and Anja Zidar. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, LChange@ACL 2022, Dublin, Ireland, May 26-27, 2022*, pages 1–10, Online. Association for Computational Linguistics.

Sebastian Nagel. 2016. CC-News.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2022. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.*, 54(1):20:1–20:39.

Terhi Nurmikko-Fuller. 2023. *Linked Data for Digital Humanities*. Routledge.

Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. Good applications for crummy entity linkers?: The case of corpus selection in digital humanities. In *Proceedings of the 13th International Conference on Semantic Systems, SEMANTiCS 2017, Amsterdam, The Netherlands, September 11-14, 2017*, pages 81–88. ACM.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? challenging transformers with generalized event knowledge. *CoRR*, abs/2107.10922.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 298–311. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. C4: Colossal clean crawled corpus.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, Kigali, Rwanda. OpenReview.net.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2021. A mathematical exploration of why language models help solve downstream tasks. In *International Conference on Learning Representations*.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Assoc. Inf. Sci. Technol.*, 60(3):538–556.

Barbara Tillett. 2002. A virtual international authority file (viaf). In *Record of a workshop on Authority Control among Chinese Korean and Japanese Languages (CJK Authority 3)*, pages 117–139.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction.

Unxos. 2013. GeoNames.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4569–4586, Online. Association for Computational Linguistics.

Donald J Waters. 2023. The emerging digital infrastructure for research in the humanities. *International Journal on Digital Libraries*, 24(2):87–102.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6397–6407, Online. Association for Computational Linguistics.

Peiran Yao, Kostyantyn Guzhva, and Denilson Barbosa. 2024. Semantic graphs for syntactic simplification: A revisit from the age of LLM. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 105–115, Bangkok, Thailand. Association for Computational Linguistics.

Gregory Yauney, Emily Reif, and David Mimno. 2023. Data similarity is not enough to explain language model performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11295–11304, Singapore. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 50–61, Online. Association for Computational Linguistics.

A The Orlando Project

The Orlando Project is an ongoing experiment in digital literary history that began in 1995. Its flagship output is a regularly updated online "textbase", *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*. As of 2023, the textbase comprises 1444 biocritical profiles of authors from 612 BCE onward, 1261 of them women, contextualized by more than 29,361 free-standing events. 2,995,455 semantic tags annotate its 9,043,111 words with structured references to 37,374 unique persons, 8,696 organizations, 12,114 place names, 47,067 titles, and 30,441 bibliographic sources, as well as embedding relationships among them.

The textbase data has been used for analysis, visualization, and interface design research; its content has fed other DH projects in women's writing; and its XML schema has served as a foundation for similar projects in the Canadian Writing Research Collaboratory[12].

Few born-digital DH resources feature such extensive annotation, since hand-annotation is costly. However, *Orlando* is representative of much DH work in being organized around profiles of significant individuals that refer to other related entities, and in using complex, nuanced language. Linking entities is a key component of DH infrastructure (Waters, 2023). More efficient and accurate EL for text such as Orlando's would provide immense benefits to DH scholars wishing to enhance their data for publication or analysis, and relationship extraction would provide even further value. Suitably packaged fine-tuned LLMs better equipped to deal with the long and elaborate sentences found in Orlando would be equally welcome by the DH community.

[12] https://cwrc.ca/

A.1 Illustrating Orlando's Complexity

The density of facts and complex sentence structures in Orlando make it a valuable DH research tool and present an interesting and potentially quite challenging dataset for NLP systems trained on simpler text.

Sentences contain lists of people with multiple parenthetical clauses and nested relations:

> "Dora Carrington formed a lively group (the Wild Group, as they were known at the Slade) with women she remained in close contact with for many years, including Dorothy Brett (later the Honourable), Barbara Hiles (later Bagenal), Ruth Humphries, and Alix Sargant-Florence (the daughter of painter Mary Sargant-Florence and later the wife of James Strachey)."

It is often ambiguous, even to a human reader, as to which relationship is referring to which entity:

> "One of her sisters and a niece, Horatia Katherine Frances Gatty (later Eden) and Christabel Maxwell, published writings about her."

There are multi-step person relations with an unnamed mother in the middle:

> "Rosina's mother's uncle, Sir John Doyle, was Lieutenant Governor of Guernsey at this time."

With a high count of meaningful clauses per sentence:

> "Louisa Baldwin's mother, a Welshwoman born Hannah Jones, was George Macdonald's second wife."

> "Her mother, born Ann Bee, died on 5 October 1766, and a widowed aunt, another Cassandra, came to keep house for the family."

B Understanding the Benchmark Dataset

Here we describe and contextualize the fields present in the Orlando (Release) JSON dataset.

Entities For each text chunk, we list entity mentions under *entities* with their *start* and *end* offsets using utf-8 encoding. For each *mention*, we include all text tagged by the annotators, as well as contextual information they added as attributes. This includes *full_name*, which for persons and organizations is a more explicit name or a reformatted name, while for places it is typically the name of the encompassing region. Person and organization mentions have manually deduplicated internal Orlando identifiers, *id*.

For each person mention, we indicate if the associated entity is the primary subject of an Orlando biography through *biography_subject*. This does not necessarily indicate the source document of a text chunk as the subject of one biography could be mentioned in another biography. It can be used as one indicator of a person's notoriety and allows for separate analyses of the writers and the people connected to their lives.

Relations We use the subject-predicate-object formation to represent extracted relations as triples. Many relations are commutative but we only explicitly list one direction. Table 8, Table 9, and Table 10 detail the relations and contextual categories present in our released dataset.

The *predicate_category* for a relation represents the high level XML tag for the text chunk while *predicate_id* and *predicate_name* represent the specific relation. The *predicate_name* is the relation, while, when available, *predicate_id* is a URI from the CWRC Ontology[13] that either exactly represents the relation or gives more specific information about the relation. For example, the *number_of_children* relation can have *predicate_id* cwrc:adoption to contextualize. Note that the same relation can be present under multiple categories, giving the relation slightly different meaning. For example *subject_studied* can have *institutional_education_context*, *self_taught_education_context*, or *domestic_education_context*.

We do provide utf-8 text spans for the objects of the triples. *object_text* contains the exact mention text of that entity, while *object_id* gives context about that entity from the annotations, when available. For places, the id is either a GeoNames URI for the place or an encompassing region, or a string listing such regions. For people and organizations, it is the de-duplicated Orlando ID. For

[13] https://sparql.cwrc.ca/

other types like occupations, id can be an identifier from sources such as the CWRC ontology or Library of Congress Subject Headings[14]. For dates, it is standard form YYYY or YYYY-MM-DD, and for nationalities, it is an ISO 3166-2 code.

C Details of Baseline Corpora

C.1 Dataset Version and Pre-processing

The English Wikipedia and Simple English Wikipedia corpora that we use in our comparisons are compiled from recent dumps: `enwiki-20230320` for ordinary English Wikipedia and `simplewiki-20230101` for Simple English.

Every corpus is pre-processed using the same pipeline as Orlando, including sentence splitting using PySBD (Sadvilkar and Neumann, 2020) followed by tokenization and entity recognition using the `en_core_web_sm` model of spaCy (Honnibal et al., 2020). For a consistent comparison across corpora, we count all entities identified by the entity recognition model of spaCy without relying on the manual entity annotations of Orlando.

C.2 Additional Lexical Complexity Statistics

We count the number of characters, tokens, and entities in each sentence and report the distributions in Figure 6 and 5 for each corpus. The lexical complexity score we use is proposed by Martin et al. (2018) which is based on the mean log-rank of word frequencies in a sentence and yields higher scores if more rare words are present in the sentence.

The distribution of sentence length in Orlando, measured by the number of tokens or characters, is similar to that of Wikipedia, with a mean higher than that of C4, CC-News, and Simple Wiki. Sentences in Orlando and Wikipedia contain more syllables and tokens, which is also reflected in the higher FKGL.

C.3 L2SCA Variables

Definitions of the 14 variables of L2SCA (Lu, 2010) are listed in Table 11. For illustration purposes, when plotting the L2SCA variables in Figure 2, we normalize the variables by min-max scaling to the range of $[0, 1]$: suppose x is the vector of raw score of a variable across all corpora, then the normalized score is

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

[14] http://id.loc.gov/authorities/subjects/

The raw values of the variables are reported in Tables 12 and 13.

D Configuration and Results of Dataset Use Cases

D.1 BLINK Entity Linking Configuration

We use the model that Wu et al. (2020) trained on a 2019 Wikipedia dump. We set the parameter k to 10, according to the authors' suggestion, so the candidate generation step selects 10 candidates, the ranking step ranks those 10, and we compute accuracy using the one highest ranked prediction. We test three options for the maximum number of contextual tokens: (1) full right and left context within the given text chunk, (2) maximum of 32 tokens on each side of the mention, and (3) maximum of 32 total context tokens as Wu et al. (2020) suggest, but the treatments all had the same results.

D.2 PURE Relation Extraction Configuration

We pre-process the chunks with PySBD (Sadvilkar and Neumann, 2020) for sentence splitting and NLTK (Bird et al., 2009) for word tokenization.

D.3 PURE Relation Extraction Example 1

PURE results are identified with square brackets and Orlando's with curly brackets.

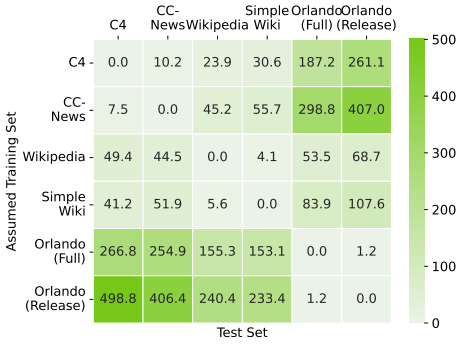
> "Coming from a {professional but not wealthy rank among the middle classes}, she seems to have had to contribute to the family income, by teaching and writing, even before [her] [parents]' deaths. A [student] at [her] [school] in [Canada] described the Walker sisters as very {English}, very dignified, and somewhat exclusive, but... excellent teachers, especially in the departments of history and English literature. Presumably she was {white} and a {Christian} —, one of her verses was appropriated as a hymn by the American {Evangelicals} Dwight L. Moody and Ira Sankey —and she may well have supported the {Temperance movement}."

D.4 PURE Relation Extraction Example 2

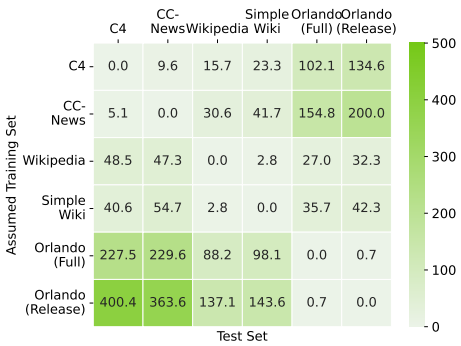
Orlando's annotations are identified with curly brackets.

> "In October 1940 he went up to {St John's College, Oxford}. He studied

{English language and literature}, and took a {first-class Honours BA} in 1943. Important friendships formed in his undergraduate days were those with {Bruce Montgomery} (who became a highly successful detective-novel writer under the name of Edmund Crispin, and dedicated one of his earliest books to Larkin) and especially the future writer {Kingsley Amis}. {Amis and Larkin constituted themselves a two-man parody factory mocking every aspect of university life: the syllabus, the dons, and the aspiring writers such as John Heath-Stubbs.}"



(a) `Llama-2-7B`



(b) `BART-large`

Figure 4: Mean squared Mahalanobis distance of the corpora in comparison to the assumed training data of LLMs.

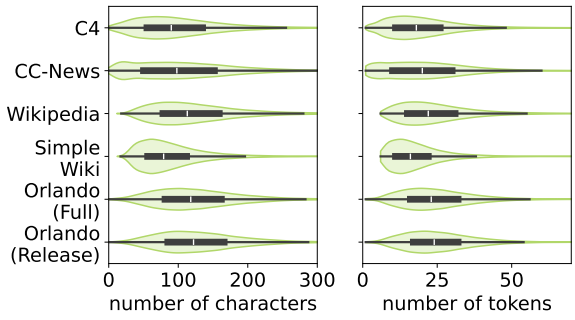


Figure 5: Distributions of the number of characters (left) and tokens (right) in sentences.

predicate_category	count
spatial_context	6818
friends_and_associates_context	5633
cultural_form_context	3275
occupation_context	3101
family_context	2236
birth_context	2210
significant_activity_context	1896
death_context	1816
institutional_education_context	1814
political_context	1676
religion_context	1006
domestic_education_context	826
intimate_relationship_context	728
social_class_context	296
self_taught_education_context	229
nationality_context	141
race_ethnicity_context	113
sexuality_context	107

Table 8: Orlando contextual categories that the relation predicates belong to, with mention counts in our Orlando (Release) dataset.

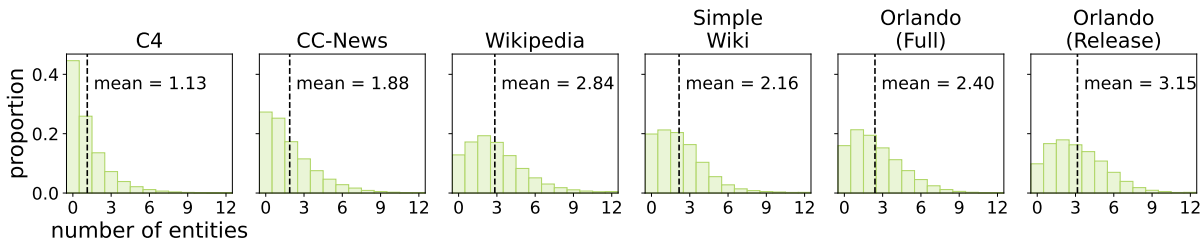


Figure 6: Distributions of the number of entities in sentences of the six corpora.

predicate_name	count	predicate_name	count
interpersonal_relationship	5700	sexuality	137
paid_occupation	2952	brother	136
travel	2336	sister	110
habitation	1846	family_based_occupation	109
relocation	1554	occupation_income	103
subject_studied	1477	son	93
religion	1256	emigration	87
social_class	1011	daughter	85
member_of	974	education_award	55
occupation	940	ethnicity	48
visit	839	cohabitant	47
date_of_birth	803	education_companion	45
place_of_birth	748	other_family	42
date_of_death	695	non_erotic_relationship	42
birth_position	659	intimate_relationship	41
place_of_death	600	contested_behaviour	40
school	594	linguistic_ability	40
nationality	545	grandfather	38
employment	540	uncle	36
activist_involvement_in	525	degree_subject	36
national_heritage	498	wife	35
husband	461	cousin	25
gendered_political_activity	424	grandmother	24
father	399	aunt	21
erotic_relationship	384	native_linguistic_ability	19
cause_of_death	361	spatial_relationship	15
volunteer_occupation	353	forebear	13
political_involvement	336	stepfather	9
mother	307	niece	9
instructor	288	grandson	8
number_of_children	262	nephew	8
political_membership	246	stepmother	7
race_colour	242	granddaughter	5
possibly_erotic_relationship	239	child	3
education_text	179	stepbrother	2
political_affiliation	175	partner	2
burial_location	160	guardian	2
degree	155	stepsister	1
migration	141	stepdaughter	1
geographic_heritage	138		

Table 9: Orlando relation predicates with the mention counts in our Orlando (Release) dataset.

predicate_name	count	predicate_name	count	predicate_name	count
birth_context		**friends_and_associates_context**		**religion_context**	
date_of_birth	803	interpersonal_relationship	5608	religion	572
place_of_birth	748	cohabitant	25	member_of	346
birth_position	659	**institutional_education_context**		social_class	28
cultural_form_context		subject_studied	736	nationality	18
social_class	768	school	585	national_heritage	11
religion	657	instructor	154	activist_involvement_in	7
nationality	434	degree	152	race_colour	4
member_of	432	education_award	52	gendered_political_activity	3
national_heritage	406	education_companion	43	geographic_heritage	3
race_colour	193	degree_subject	36	political_affiliation	3
geographic_heritage	106	education_text	29	sexuality	3
political_affiliation	44	contested_behaviour	27	ethnicity	2
political_involvement	41	**intimate_relationship_context**		linguistic_ability	2
sexuality	39	erotic_relationship	384	political_involvement	2
ethnicity	35	possibly_erotic_relationship	239	political_membership	2
activist_involvement_in	29	non_erotic_relationship	42	**self_taught_education_context**	
linguistic_ability	28	intimate_relationship	41	subject_studied	174
political_membership	27	cohabitant	22	education_text	32
gendered_political_activity	23	**nationality_context**		instructor	15
native_linguistic_ability	13	nationality	50	education_award	3
death_context		national_heritage	46	contested_behaviour	2
date_of_death	695	social_class	16	degree	2
place_of_death	600	geographic_heritage	8	school	1
cause_of_death	361	religion	8	**sexuality_context**	
burial_location	160	race_colour	5	sexuality	93
domestic_education_context		member_of	4	activist_involvement_in	3
subject_studied	567	ethnicity	2	gendered_political_activity	3
instructor	119	native_linguistic_ability	2	social_class	3
education_text	118	**occupation_context**		nationality	1
contested_behaviour	11	paid_occupation	1560	political_affiliation	1
school	8	employment	540	political_membership	1
education_companion	2	occupation	520	race_colour	1
degree	1	volunteer_occupation	314	religion	1
family_context		occupation_income	103	**significant_activity_context**	
husband	461	family_based_occupation	64	paid_occupation	1392
father	399	**political_context**		occupation	420
mother	307	activist_involvement_in	481	family_based_occupation	45
number_of_children	262	gendered_political_activity	391	volunteer_occupation	39
brother	136	political_involvement	289	**social_class_context**	
sister	110	political_membership	213	social_class	183
son	93	member_of	177	nationality	32
interpersonal_relationship	92	political_affiliation	123	national_heritage	20
daughter	85	religion	2	member_of	12
other_family	42	race_colour	29	race_colour	10
grandfather	38	national_heritage	15	geographic_heritage	9
uncle	36	social_class	13	religion	9
wife	35	geographic_heritage	12	activist_involvement_in	4
cousin	25	nationality	10	gendered_political_activity	4
grandmother	24	linguistic_ability	9	political_affiliation	4
aunt	21	ethnicity	8	political_involvement	3
forebear	13	religion	7	political_membership	3
niece	9	native_linguistic_ability	4	ethnicity	1
stepfather	9	member_of	3	linguistic_ability	1
grandson	8	activist_involvement_in	1	sexuality	1
nephew	8	political_involvement	1	**spatial_context**	
stepmother	7	sexuality	1	travel	2336
granddaughter	5			habitation	1846
child	3			relocation	1554
guardian	2			visit	839
partner	2			migration	141
stepbrother	2			emigration	87
stepdaughter	1			spatial_relationship	15
stepsister	1				

Table 10: Orlando relation predicates with mention counts in our Orlando (Release) dataset. Predicates are repeated in each contextual category (bolded text) in which they appear.

Code	Measure	Definition
MLC	Mean length of clause	# of words / # of clauses
MLS	Mean length of sentence	# of words / # of sentences
MLT	Mean length of T-unit	# of words / # of T-units
C/S	Sentence complexity ratio	# of clauses / # of sentences
C/T	T-unit complexity ratio	# of clauses / # of T-units
CT/T	Complex T-unit ratio	# of complex T-units / # of T-units
DC/C	Dependent clause ratio	# of dependent clauses / # of clauses
DC/T	Dependent clauses per T-unit	# of dependent clauses / # of T-units
CP/C	Coordinate phrases per clause	# of coordinate phrases / # of clauses
CP/T	Coordinate phrases per T-unit	# of coordinate phrases / # of T-units
T/S	Sentence coordination ratio	# of T-units / # of sentences
CN/C	Complex nominals per clause	# of complex nominals / # of clauses
CN/T	Complex nominals per T-unit	# of complex nominals / # of T-units
VP/T	Verb phrases per T-unit	# of verb phrases / # of T-units

Table 11: Descriptions and definitions of variables of L2SCA. The code is used in Figure 2. The table is adapted from Lu, 2010 (pp. 479).

Corpus	MLC	MLS	MLT	C/S	C/T	CT/T	DC/C
C4	11.4955	17.7755	16.8636	1.5463	1.467	0.3563	0.3292
CC-News	11.4438	21.5108	19.6258	1.8797	1.715	0.4279	0.3662
Wikipedia	14.5312	21.079	19.6803	1.4506	1.3543	0.2726	0.2524
Simple Wiki	11.8485	15.5475	14.9764	1.3122	1.264	0.2092	0.2056
Orlando (Full)	13.8285	24.5383	21.1579	1.7745	1.53	0.4033	0.3378
Orlando (Release)	16.0128	26.7895	23.6033	1.673	1.474	0.3696	0.3138

Table 12: Raw scores of the first seven L2SCA variables of the six corpora.

Corpus	DC/T	CP/C	CP/T	T/S	CN/C	CN/T	VP/T
C4	0.4829	0.305	0.4474	1.0541	1.2147	1.7819	1.9977
CC-News	0.6281	0.2355	0.4039	1.096	1.2628	2.1656	2.2858
Wikipedia	0.3419	0.3694	0.5003	1.0711	1.6486	2.2327	1.6942
Simple Wiki	0.2599	0.2829	0.3576	1.0381	1.3151	1.6622	1.4589
Orlando (Full)	0.5169	0.3188	0.4877	1.1598	1.3471	2.0611	1.9369
Orlando (Release)	0.4625	0.3607	0.5316	1.135	1.4648	2.1592	1.8236

Table 13: Continuation of Table 12. Raw scores of the rest of L2SCA variables of the six corpora.