1 Introduction

Much work involving the computational analysis of text in the humanities draws on methods and tools in natural language processing (NLP)—an area of research focused on reasoning about the linguistic structure of text. Over the past 70 years, the ways in which we process natural language have blossomed into dozens of applications, moving far beyond early work in speech recognition (Davis et al., 1952) and machine translation (Bar-Hillel, 1960) to include question answering, summarization, and text generation.

While the direction of many avenues of research in NLP has tended toward the development of end-to-end-systems—where, for example, a model learns to solve the problem of information extraction without relying on intermediary steps of part-of-speech tagging or syntactic parsing—one area where the individual low-level linguistic representations are important comes in analyses that depend on them for argumentation. Work in the computational humanities, in particular, has drawn on this representation of linguistic structure for a wide range of work. Automatic part-of-speech taggers have been used to explore poetic enjambment (Houston, 2014) and characterize the patterns that distinguish the literary canon from the archive (Algee-Hewitt et al., 2016). Syntactic parsers have been used to attribute events to the characters who participate in them (Jockers and Kirilloff, 2016; Underwood et al., 2018) and characterize the complexity of sentences in Henry James (Reeve, 2017). Coreference resolution has been used to explore the prominence of major and minor characters as a function of their gender (Kraicer and Piper, 2018). Named entity recognizers have been used to explore the relationship between places in British fiction and cultural identity (Evans and Wilkens, 2018); geographic markers extracted from NER have been used to create visualizations of the places mentioned in texts, both for toponyms in Joyce’s *Ulysses* (Derven et al., 2014) and short fiction by Edward P. Jones (Ramsby and Ossom-Williamson, 2019). Topics help organize a range of work in the humanities, from identifying the characteristics of colonial fiction in Australian newspapers (Bode, 2018) to surfacing editorial labor in 19th-century US ones (Klein, 2020). And moving beyond text to sound studies, work has also explored using NLP to extract prosodic features from texts (Clement et al., 2013) and directly model audio data to investigate questions revolving around applause (Clement and McLaughlin, 2018) and poet voice (MacArthur et al., 2018). In each of these cases, the fundamental research question is not in solving an NLP problem, but in treating NLP as an algorithmic measuring device—representing text in a way that allows a comparison of measures to be made, whether for the purpose of explicit hypothesis testing or exploratory analysis.

The demands of literary and historical texts have certainly influenced the design of NLP over its lifetime: some of the earliest work on text generation, including Novel Writer (Klein et al., 1973) and TaleSpin (Meehan, 1977), were both designed to explicitly generate narrative stories; the field of authorship attribution, which was first proposed by Mendenhall (1887) to discriminate the works of Francis Bacon, Shakespeare

---

and Christopher Marlowe, later drove the pioneering work on the Federalist Papers by Mosteller and Wallace (1964) and is now used in applications as far removed as forensic analysis; and BERT (Devlin et al., 2019), one of the most widely used models in the family of contextual representation learning methods responsible for many of the recent advancements in NLP, is trained not only on English Wikipedia, but also on the BookCorpus, a collection of 11,038 books from the self-publishing platform smashwords.com—attesting again to the wealth of commonsense knowledge that fiction can provide.

But more broadly, mainstream NLP has tended to focus on a relatively small set of domains—including news, which forms the overwhelming basis for benchmark corpora including MUC (Sundheim, 1991), the Penn Treebank (Marcus et al., 1993), ACE (Walker et al., 2006), the New York Times Annotated Corpus (Sandhaus, 2008), and OntoNotes (Hovy et al., 2006); and Wikipedia, which provides the benchmark datasets for question answering (Rajpurkar et al., 2016, 2018) and named entity linking (Cucerzan, 2007), and has provided the training material for many language models in multiple languages (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2019).

In many ways, however, literary texts push the limits of methods in natural language processing that have been optimized for these narrow domains—the long, complex sentences in novels strain the limits of syntactic parsers with super-linear computational complexity, their use of figurative language challenges representations of meaning based on neo-Davidsonian semantics, and their long length rules out existing solutions for problems like coreference resolution that expect short documents with a small set of candidate antecedents. If methods in NLP are to be used for analyses to help drive humanistic insight, this disparity necessitates developing resources and models in NLP that are born-literary—models trained on specifically literary data that attests to the phenomena that we might want to measure, that encodes biases we deem more appropriate than those encoded in other datasets designed for other purposes (such as identifying acts of terrorism in MUC), and that specifically considers the needs of researchers working with fictional and historical material—for example, by historicizing categories of gender (Mandell, 2019). A focus on born-literary NLP not only seeks to improve the state of the art for NLP in the domain of literature, but also examines the specific research questions that are only afforded within literature—while other work in NLP touches on aspects of narrative that are common with news (such as inferring narrative event chains (Chambers, 2011)), literature presents a number of unique opportunities, including modeling suspense, the passage of time, focalization, and more. Importantly, born-literary NLP also entails a widening of opportunity—while much work in NLP has been dominated by researchers in the fields of computer science, information science, and linguistics, the researchers poised to making the biggest advances in this area are those with training in the humanities—who can not only leverage expertise in the specific subject matter to define the appropriate boundaries of datasets, but who can also use their own disciplinary expertise to define the literary phenomena that are worth modeling in the first place.

2 Performance across domains

Progress in natural language processing is primarily driven by comparative performance on benchmark datasets—progress in phrase-structure syntactic parsing, for example, has been defined for 30 years by performance on the Penn Treebank (Marcus et al., 1993). Benchmark datasets provide an external control: given a fixed dataset (ideally with pre-determined train and test partitions), researchers can be more confident that an increase in performance by their model for the task relative to another model can be attributed to their work alone, and not simply be due to incomparable performance on different datasets.

At the same time, benchmark datasets tend to myopically focus attention on the domains they represent, and generalization performance beyond those datasets can be quite poor. Table 1 represents a metareview illustrating this performance degradation across a range of training/test scenarios. A model trained on one domain may yield high performance when evaluated on data from that same domain, but often suffers a steep
drop in performance when evaluated on data from another domain. In the few cases that involve training on news and evaluating on literature, these drops in performance can amount to 20 absolute points or more, effectively rendering a tool unusable.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Task</th>
<th>In domain</th>
<th>Acc.</th>
<th>Out domain</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rayson et al. (2007)</td>
<td>POS</td>
<td>English news</td>
<td>97.0%</td>
<td>Shakespeare</td>
<td>81.9%</td>
</tr>
<tr>
<td>Scheible et al. (2011)</td>
<td>POS</td>
<td>German news</td>
<td>97.0%</td>
<td>Early Modern German</td>
<td>69.6%</td>
</tr>
<tr>
<td>Moon and Balbridge (2007)</td>
<td>POS</td>
<td>WSJ</td>
<td>97.3%</td>
<td>Middle English</td>
<td>56.2%</td>
</tr>
<tr>
<td>Pennacchiotti and Zanzotto (2008)</td>
<td>POS</td>
<td>Italian news</td>
<td>97.0%</td>
<td>Dante</td>
<td>75.0%</td>
</tr>
<tr>
<td>Derczynski et al. (2013b)</td>
<td>POS</td>
<td>WSJ</td>
<td>97.3%</td>
<td>Twitter</td>
<td>73.7%</td>
</tr>
<tr>
<td>Gildea (2001)</td>
<td>PS parsing</td>
<td>WSJ</td>
<td>86.3 F</td>
<td>Brown corpus</td>
<td>80.6 F</td>
</tr>
<tr>
<td>Lease and Charniak (2005)</td>
<td>PS parsing</td>
<td>WSJ</td>
<td>89.5 F</td>
<td>GENIA medical texts</td>
<td>76.3 F</td>
</tr>
<tr>
<td>Burga et al. (2013)</td>
<td>Dep. parsing</td>
<td>WSJ</td>
<td>88.2%</td>
<td>Patent data</td>
<td>79.6%</td>
</tr>
<tr>
<td>Pekar et al. (2014)</td>
<td>Dep. parsing</td>
<td>WSJ</td>
<td>86.9%</td>
<td>Broadcast news</td>
<td>79.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Magazines</td>
<td>77.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Broadcast conversation</td>
<td>73.4%</td>
</tr>
<tr>
<td>Derczynski et al. (2013a)</td>
<td>NER</td>
<td>CoNLL 2003</td>
<td>89.0 F</td>
<td>Twitter</td>
<td>41.0 F</td>
</tr>
<tr>
<td>Bamman et al. (2019)</td>
<td>Nested NER</td>
<td>News</td>
<td>68.8 F</td>
<td>English literature</td>
<td>45.7 F</td>
</tr>
<tr>
<td>Bamman et al. (2020)</td>
<td>Coreference</td>
<td>News</td>
<td>83.2 F</td>
<td>English literature</td>
<td>72.9 F</td>
</tr>
<tr>
<td>Naik and Rose (2020)</td>
<td>Events</td>
<td>News</td>
<td>82.6 F</td>
<td>English literature</td>
<td>44.1 F</td>
</tr>
</tbody>
</table>

Table 1: In-domain and out-of-domain performance for several NLP tasks, including POS tagging, phrase structure (PS) parsing, dependency parsing, named entity recognition, coreference resolution and event trigger identification. Accuracies are reported in percentages; phrase structure parsing, NER, coreference resolution and event identification are reported in F1 measure.

Perhaps more pernicious than a simple drop in performance, however, are the forms of representational bias that are present in any dataset. As Bamman et al. (2019) point out for literary texts, an entity recognition model trained on news (the ACE 2005 dataset) is heavily biased toward recognizing men, simply given the frequency with which men are present in that news data; when tested on literature, where men and women are mentioned with greater parity, the recall at recognizing women is disparately poor, recognizing only 38.0% of mentions, compared to 49.6% for men (a difference of −11.6 points). A model trained natively on literature, however, corrects this disparity, recognizing 69.3% of mentions who are men and 68.2% of those who are men (a difference +1.1 points).

One motivation for born-literary NLP is to simply improve this dismal performance—if a model is able to reach an F-score of 68.8 for entity recognition in English news, then we should not have to settle for an F-score of 45.7 for English literature. But beyond that overall goal is a concern that strikes at the heart of methods in the computational humanities—if we are using empirical methods as algorithmic measuring devices, then absolute accuracy is less important than the source of any measurement error: if error is non-random, such that measurement accuracy is dependent on a variable that is at the core of subsequent analysis (such as gender), then we need to account for it. While methods from the social sciences that deal with survey data—like multilevel regression and poststratification (Gelman and Little, 1997)—may provide one means of correcting the disparity between a biased sample and the true population they are meant to reflect (if error rates are known and we only care about aggregate statistics), there are many situations where such methods fail. An alternative is much simpler: we can train models natively on literary data, and encode the biases present in representatives of the data we will later analyze.
3 LitBank

By training a model on data that resembles what it will see in the future, we can expect our performance on that future data to be similar to the performance on data we’ve seen during training. Several efforts have done just this for a range of linguistic phenomena, including part-of-speech tagging (Mueller, 2015) and syntactic parsing in a variety of historical and literary registers—including Portuguese (Galves and Faria, 2010), Greek and Latin (Bamman and Crane, 2011; Passarotti, 2007; Haug and Jøhndal, 2008) and English (Taylor and Kroch, 2000; Kroch et al., 2004; Taylor et al., 2006). By annotating data within the domain we care to analyze, we can train better methods to analyze data that looks similar to it in the future.

LitBank is one such resource: an open-source, born-literary dataset to support a variety of contemporary work in the computational humanities working with English texts. To date, it contains 210,532 tokens drawn from 100 different English-language novels, annotated for four primary phenomena: entities, coreference, quotations and events. By layering multiple phenomena on the same fixed set of texts, the annotations in LitBank are able to support interdependence between the layers—coreference, for example, groups mentions of entities (Tom, the boy) into the unique characters they refer to (TOM SAWYER), and quotation attribution assigns each act of dialogue to the unique character (i.e., coreference chain) who speaks it.

3.1 Sources

The texts in LitBank are all drawn from public domain texts in Project Gutenberg, and include a mix of high literary style (e.g., Edith Wharton’s Age of Innocence, James Joyce’s Ulysses) and popular pulp fiction (e.g., H. Rider Haggard’s King Solomon’s Mines, Horatio Alger’s Ragged Dick). All of the texts in LitBank were originally published before 1923, and, as figure 1 illustrates, predominantly fall at the turn of the 20th century.

3.2 Phenomena

Entities. Entities define one of the core objects of interest in the computational humanities; entities capture the characters that are involved in stories, the places where they operate, and the things they interact with. Much work in the computational humanities reasons about these entities, including character (Underwood et al., 2018), places (Evans and Wilkens, 2018) and objects (Tenen, 2018), and has focused on improving NER for specifically the literary domain (Brooke et al., 2016).

Traditional NER systems for other domains like news typically disallow hierarchical structure within names—flat structure is easier to reason about computationally (where it can be treated as a single-layer sequence labeling problem) and largely fits the structure of the phenomenon, where common geo-political entities (like Russia) and people (Bill Clinton) lack hierarchical structure. But literature abounds with hierarchical entities, many of which are not named at all:

```
PER
PER
PER

... the elder brother of Isabella’s husband
```
In this passage from Jane Austen’s *Emma*, we see multiple entities expressed: *Isabella*, *Isabella’s husband*, and *the elder brother of Isabella’s husband*. Even though they are not named, all are potentially significant as mentions of characters within this story.

To capture this distinctive characteristic of literary texts, the first annotation layer of LitBank (described in Bamman et al. (2019) and Bamman et al. (2020)) identifies all entities of six types—people (PER), facilities (FAC), geo-political entities (GPE), locations (LOC), organizations (ORG) and vehicles (VEH) and classifies their status as a proper name (PROP), common noun phrase (NOM), or pronoun (PRON). As table 3 illustrates, the proportion of entities that traditional NER would capture (PROP) is quite small—common entities (her sister) are mentioned nearly three times as often as proper names (Jane), and both far less frequently than pronouns.

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>24,180</td>
<td>83.1%</td>
</tr>
<tr>
<td>FAC</td>
<td>2,330</td>
<td>8.0%</td>
</tr>
<tr>
<td>LOC</td>
<td>1,289</td>
<td>4.4%</td>
</tr>
<tr>
<td>GPE</td>
<td>948</td>
<td>3.3%</td>
</tr>
<tr>
<td>VEH</td>
<td>207</td>
<td>0.7%</td>
</tr>
<tr>
<td>ORG</td>
<td>149</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

Table 2: Counts of entity type.

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRON</td>
<td>15,816</td>
<td>54.3%</td>
</tr>
<tr>
<td>NOM</td>
<td>9,737</td>
<td>33.5%</td>
</tr>
<tr>
<td>PROP</td>
<td>3,550</td>
<td>12.2%</td>
</tr>
</tbody>
</table>

Table 3: Counts of entity category.

What can we do with books labeled with these entity categories? At their simplest, entities provide an organizing system for the collection, as Wolfe (2019) has demonstrated by applying models trained on LitBank to texts in the Black Books Interactive Project ([https://bbip.ku.edu](https://bbip.ku.edu))—simply providing a ranked list of the most frequent people and places mentioned in a text provides a high-level overview of the content of a work.

At the same time, entity types abstract away common patterns that provide insight into narrative structure. As McClure (2017) points out at the scale of individual words, many terms exhibit strong temporal associations with the narrative time of a book: significant plot elements like death show up near the end of novel, while many terms that introduce people show up earlier. By examining the broad trends with which entire entity categories—like people—are mentioned over the scale of an entire novel, and further distinguishing between proper name mentions, common noun phrase mentions, and pronouns, we can see (fig. 2) that different temporal dynamics influence each one: while proper names and pronouns increase in frequency as a book progresses from its beginning to end, common noun phrases such as “the boy” show a marked decline in frequency.

Figure 2: Distribution over narrative time of automatically predicted PROP, NOM and PRON person entities in 100 English-language books from Project Gutenberg, excluding all paratextual material from Project Gutenberg (legal boilerplate) and the print original (tables of contents, advertisements, etc.).
Coreference. Identifying the spans of text that correspond to entity classes is useful for aggregating together those mentions whose names are determinative of their identity—for example, just about every mention of **New York City** in a text will refer to the unique city known by that name. But there is often ambiguity in mapping mentions to entities—e.g., for the 72 mentions of “Miss Bennett” in *Pride and Prejudice*, which specific mentions refer to Elizabeth, Jane, Lydia, Mary, or Kitty? This issue is exacerbated when pronouns are also considered as potential entities (who does she refer to?). Coreference resolution is the challenge of clustering these ambiguous mentions so that all mentions that co-refer to the same distinct entity are placed in the same cluster.

The benchmark dataset for coreference in English is OntoNotes 5 (Weischedel et al., 2012), which includes the domains of news (broadcast, magazine, and newswire), conversation, the web, and even some fiction (though restricted to include only the Bible). There are many ways, however, in which coreference in literature differs from that in factual textual sources, including the delayed revelation of identity (common in detective stories and mysteries, for example), in which two characters portrayed as separate entities are revealed to be the same person. Narratives in literary texts also tend to span longer time frames than news articles—perhaps years, decades, or even centuries—which raises difficult questions on the metaphysical nature of identity (e.g., is Shakespeare’s **LONDON** of 1599 the same entity as Zadie Smith’s **LONDON** in the year 2000?).

To address these issues, the second layer of annotations in LitBank (described in Bamman et al. (2020)) covers coreference for the entities annotated above; we specifically consider the ways in which literary coreference differs from coreference in news and other short, factual texts, and manually assign the 29,103 mentions annotated above into 7,235 unique entities. As a result, coreference models trained on this native literary data perform much better on literary text (79.3F average F1 score) than those trained on OntoNotes (average 72.9 F). This joins existing datasets for literature, including the work of Vala et al. (2015), which annotates character aliases in *Sherlock Holmes, Pride and Prejudice* and *The Moonstone*, and annotated datasets of conference in German novels (Krug et al., 2017) and plays (Pagel and Reiter, 2020).

What does coreference make possible for cultural analysis? Coreference is critical for aggregating information about distinct entities like characters—Underwood et al. (2018), for example, measures the amount of “attention” that male and female characters receive in novels over 200 years of literary history by counting up the number of actions each character participates in; this is only possible by having a stable entity for each character (given through coreference) that such counts can apply to; and given that over half of all entity mentions are pronominal in nature means that including pronouns in coreference is important for that characterization. Less work has explored the potential of coreference for other entity categories beyond people, but coreference for other classes such as places—which include natural locations like **the marsh** and **the forest** (LOC), human-created structures like houses and rooms (FAC) and geo-political entities like cities, countries and villages (GPE)—is in many ways a precondition for the analysis of setting and its relationship to plot; in order to chart that characters in *Lord of the Rings* begin in *The Shire*, venture to **Mount Doom** and return **home** in the end, we need to understand that **The Shire** and **home** refer to the same physical location.

Quotation attribution. Much work in the computational humanities has explored the affordances of speaker attribution—identifying the speaker of a given piece of dialogue. Such attributed dialogue has been used in the past to create character networks by defining characters to be nodes and forming an edge between two characters who speak to one another (Elson et al., 2010). Born-literary quotation data exists for both English and German: for English, this data encompasses Austen’s *Pride and Prejudice* and *Emma* as well as Chekhov’s *The Steppe* (He et al., 2013; Muzny et al., 2017b), while the Columbia Quoted Speech Corpus (Elson and McKeown, 2010) includes six texts by Austen, Dickens, Flaubert, Doyle and Chekhov. For German, Brunner et al. (2020) annotate 489,459 tokens for speech, thought and writing, including direct, indirect, free indirect and reported speech.
In order to provide a more diverse set of data for English, the third layer of LitBank (described in Sims and Bamman (2020)) includes dialogue attribution for the 100 texts in its collection. This includes 1,765 dialogue acts across 279 characters in the 100 novels present in LitBank, and allows us to measure the accuracy of both quotation identification and attribution across a much broader range of texts than previously studied. Table 4 provides a summary of the characters with the most dialogue in this annotated dataset.

What can we do with such attribution data? Expanding on the use of quotations to define character interaction networks, Sims and Bamman (2020) use quotations to extract atomic units of information and measure how those units propagate through the network defined by people speaking to each other (finding both that information propagates through weak ties and that women are often depicted as being the linchpins of information flow).

Quotation also enables direct analysis of character idiolects, allowing us to ask what linguistic properties differentiate dialogue from narrative (Muzny et al., 2017a) and the speech of characters from each other (Vishnubhotla et al., 2019)—including the degree to which speech is predictive of other traits like personality (Flekova and Gurevych, 2015). While Sims and Bamman (2020) exploit the notion of “listeners” of dialogue in order to track propagation, there is a range of work to be done in analyzing what differentiates the speech of a single character as they address different listeners, following the fundamental principle of audience design (Bell, 1984).

### Events

While entity recognition captures the important characters and objects in literature, recognizing events is important for grounding actions in plot. Event-annotated datasets in NLP have historically focused on the domain of news, including MUC (Sundheim, 1991), ACE (Walker et al., 2006) and DEFT (Aguilar et al., 2014), with some exceptions—Sprugnoli and Tonelli (2017), in particular, present an annotated dataset for historical texts that captures important classes of events in consultation with historians. But the depiction of events in literary texts tend to be very different from events in news—literary texts include long, complex structures of narrative, and multiple diachronic frames (in which some events properly belong to the space of the plot, while others exist only in the commentary by the author). To address the specificity of the problem for literature, the fourth layer of annotation in LitBank (described in Sims et al. (2019)) focuses on realis events—events that are depicted as actually taking place (not hypotheticals, conditionals, or extra-diachronic events). The criteria for what constitutes a realis event falls into four distinct categories (in all examples below, all and only realis events appear in boldface):

- **Polarity**: events must be asserted as actually occurring, and not marked as having *not* taken place (John walked by Frank and didn’t say hello).
- **Tense**: event must be in past or present tense, not future events that have not yet occurred (John walked to the store and will buy some groceries).
- **Specificity**: events must involve specific entities and take place at a specific place and time (John walked to work Friday morning) and not unqualified statements about classes (Doctors walk to work).
- **Modality**: events must be asserted as actually occurring, as distinct from events that are the targets of other modalities, including beliefs, hypotheticals, desires, etc. (John walked to the store to buy some groceries).

We annotate a total of 7,849 events in the 100 novels of LitBank. As Naik and Rose (2020) have shown, models trained natively on news (TimeBank) tend to perform quite poorly on LitBank (leading to a cross-domain drop in performance of 38.5 points), attesting to the benefit of annotated data within the domain we

<table>
<thead>
<tr>
<th>Character</th>
<th>Text</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buck Mulligan</td>
<td>Ulysses</td>
<td>43</td>
</tr>
<tr>
<td>Convict</td>
<td>Great Expectation</td>
<td>33</td>
</tr>
<tr>
<td>Mrs. Bennett</td>
<td>Pride and Prejudice</td>
<td>29</td>
</tr>
<tr>
<td>Ragged Dick</td>
<td>Ragged Dick</td>
<td>28</td>
</tr>
<tr>
<td>Mr. Bennett</td>
<td>Pride and Prejudice</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 4: Characters with most annotated dialogue.
care about.

What can we do with events? In Sims et al. (2019), we show that examining realis events shows a meaningful difference between popular texts and texts with high prestige (marked as the number times an author’s works were reviewed by elite literary journals, following Underwood (2019)). Authors with high prestige not only present a lower intensity of realis events in their work than authors of popular texts, but also have much more variability in their rates of eventfulness; popular texts, in contrast, have much less freedom in this respect, exhibiting a much narrower range of variation. Additionally, Sap et al. (2020) builds on this work by leveraging models trained on LitBank events to measure the difference between imagined and recalled events, showing that stories that are recalled contain more realis events than those that are entirely fictional. While existing work to date has focused on measurements of events on their own, there is much space for exploring the interaction between events and other narrative components—including characters (which characters participate in the most realis events?) and pacing (which works have the highest ratio of realis events per page?).

3.3 Coverage

One critique we might level at this work is that “literature” is not a monolithic domain—and, in fact, the differences between individual texts that fall into what we call literature can be much greater than the cross-domain difference between a random novel and a news article. One of the biggest differences on this front is due to time—methods that are trained on texts published before 1923 will help us little in recognizing entities in contemporary novels like Facebook, the jet and Tesla and events like googling and texting.

LitBank contains texts published before 1923 in order to work exclusively with public domain texts, so that the original text can be published along with the annotations we layer. While texts published before 1923 capture a wide range of literature, this decision is restrictive, missing nearly a century of more contemporary texts, along with the more diverse voices represented in novels published today. Our current efforts are focused on expanding LitBank to include samples from 500 books published between 1924–2020, including 100 works written by Black authors drawn from the Black Books Interactive Project, 100 works by global Anglophone writers, 100 bestsellers, 100 prizewinning books, and 100 works of genre fiction. While these texts are in copyright, we will publish samples of the texts along with linguistic annotations in order to enable reproducibility under the belief that doing so is a transformative use of the original text that adds new value and does not effect the market for the original work, and hence falls under the protections of fair use (Samberg and Hennesy, 2019).

At the same time, LitBank also is focused on works of fiction in the English language, further exacerbating what Roopika Risam notes is “the Anglophone focus of the field” of digital humanities (Risam, 2016); in many cases, components of the NLP pipeline that work reasonably well for English perform quite poorly for other languages, such as NER for Spanish literary texts (Isasi, 2017). Current work is also focused on expanding the languages represented in LitBank to include Chinese and Russian, with more to follow.

4 Born-literary questions

There is a rich research space building methods and datasets to adapt existing components of the NLP pipeline to work better on literary texts. But at the same time, an emphasis on born-literary NLP requires attending to specifically literary questions that current NLP systems cannot directly address. As Lauren Klein notes, we should not let our questions be guided by the performance of our algorithms (Klein, 2018). What are these questions that are uniquely literary?

One set of questions models the relationship between speakers and the texts they read, including the state of knowledge that we might surmise a reader has at a given point in the text. This is a problem that
uniquely pertains to narrative text, where a reader builds up a model of the represented world over the course of reading, and has access to facts that obtain within that world and predictions that they might make about future events within it. While some work in NLP addresses the question of the temporal order with which stories unfold (Mostafazadeh et al., 2016), one phenomenon that is uniquely literary is suspense—the potential anxious uncertainty about what is yet to come. Algee-Hewitt (2016) models this phenomenon by soliciting judgments of suspense from readers and building a model to predict that rating from an input passage that is 2% the length of a book, and Wilmot and Keller (2020) model suspense in short stories by measuring the reduction in future uncertainty. While most document-level classification tasks presume simultaneous access to the entirety of a text, suspense is one phenomenon where the sequential ordering of narrative is critical for understanding—we are essentially modeling a reader’s state of mind at time $t$ having read the text through time $t$ but not after it. Recent work in the computational humanities has begun to explore this phenomenon from the perspective of intratextual novelty and repetition (McGrath et al., 2018; Long et al., 2018)—modeling the degree to which authors repeat information within a book—but there are many other related phenomena (such as foreshadowing) that remain to be explored.

A second set of questions arises due to the formal nature of novels and longer literary texts—unlike news, Wikipedia articles, and tweets, novels are long (roughly 100,000 word long on average). This length presents challenges for NLP that was designed for other domains—in particular, interesting questions that we might ask of the nature of objects and things more generally (Brown, 2001) are resisted by the quality of coreference resolution for common entities like cars, guns, and houses over long distances of narrative time. Tenen (2018) is one example of the kind of work that can done when reasoning about the nature of objecthood—in that case, considering the density of objects mentioned. What we often want is not only a measure of how objects in the abstract behave, but how specific objects are depicted—such as the eponymous houses in Forster’s Howards End, Hawthorne’s House of Seven Gables or Danielewski’s House of Leaves. Characterizing those distinct houses requires us to identify when any individual mention of the word house refers to the named house in question—a task challenging even for short documents, but far more difficult at the moment for hundreds of mentions of such a common phrase potentially describing dozens of unique entities. Even though this is more of a computational challenge than a literary one, it is one driven exclusively by the characteristics of literary texts, and is unlikely to be solved by anyone not working in the computational humanities.

Finally, a third set of questions are narratological ones—how do we recognize the individual components of narrative, and assemble them together into a representation of plot? A wealth of work has explored this question from different angles, including inferring sentiment arcs (Jockers, 2015; Reagan et al., 2016), identifying “turning points” in movies (Papalampidi et al., 2019), disentangling storylines in Infinite Jest (Wallace, 2012), and segments in The Waste Land (Brooke et al., 2012), identifying Proppian narrative functions in fairy tales (Finlayson, 2015, 2016) and modeling free indirect speech (Brunner et al., 2019) and stream of consciousness (Long and So, 2016), and measuring the passage of time (Underwood, 2016). Much of the difficulty for modeling complex narratological phenomena is embedded in the difficulty of simply operationalizing what a concept like “plot” means as a computational form. Recent work attempts to tackle this theoretical question head on, by comparing different narratological annotation schemes as a first step toward computational modeling (Reiter et al., 2019). But in many ways, modeling narratological questions is uniquely positioned at the intersection of computation and the humanities—requiring not only expertise in models of linguistic structure but also a deep foundation in literary and narrative theory (Genette, 1982, 1983; Bal, 2017). The breadth of areas in this space—ranging from identifying characters and settings to inferring storylines and hierarchical narrative levels—makes modeling narratological phenomena one of the most vibrant areas poised for transformative work going forward.
5 Future

There is a range of work in the computational humanities that relies on linguistic structure—established phenomena like named entity recognition, uniquely literary tasks like predicting the passage of time, and a variety of opportunities on the horizon—that raise the potential to generate insight by considering the inherent structure present within text. While the field of natural language processing has focused for years on developing the core computational infrastructure to infer linguistic structure, much work remains to both adapt those methods to the domain(s) of literature, and also to explore the unique affordances that literature provides for computational inquiry. For existing tasks—entity recognition, coreference resolution, event identification, quotation attribution—one straightforward solution exists: we need to create more annotated data comprised of the literary texts that form the basis of our analyses, for both training (to improve the models on this domain) and evaluation (so that we know they work). LitBank provides one such resource; while this dataset is expanding to encompass a greater variety of texts, it will always hold gaps—both in its representation and in the phenomena it contains; more annotated data is always needed.

Annotated data from literary texts provides a solution to one issue in born-literary NLP; how do we go about tackling new born-literary questions, including those research areas outlined above? For several components of these problems, we can fall back on time-tested strategies: if we can operationalize a concept and annotate its presence in text to a reliable degree, we can annotate texts and train models to predict those human judgments for new texts we haven’t labeled yet. The complexity of modeling can range from straightforward sentence-level classification problems of suspense to complex hierarchical models of narrative levels; while the design of some models will require training in NLP, the most important parts of this work are often outside the realm of computation—including the insight into theory that can provide a scaffolding for an empirical method, the ability to circumscribe the boundaries of a problem that are feasible enough to address with computational methods while also being rich enough to sustain their relevance for humanistic inquiry, and the creativity needed to identify the questions worth asking in the first place. Like its broader field of the computational humanities, born-literary NLP necessarily draws on expertise in both disciplines that comprise its community of practice.

Acknowledgments

The research reported in this article was supported by an Amazon Research Award and NSF CAREER grant IIS-1942591, along with resources provided by NVIDIA and Berkeley Research Computing.

References


