LitBank: Born-Literary Natural Language Processing*

David Bamman
University of California, Berkeley
dbamman@berkeley.edu

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 (Rambsy 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

*To appear as David Bamman (2020), "LitBank: Born-Literary Natural Language Processing," in: Jessica Marie Johnson, David Mimno, and Lauren Tilton (eds.), *Computational Humanities*, Debates in Digital Humanities.

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.

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

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.

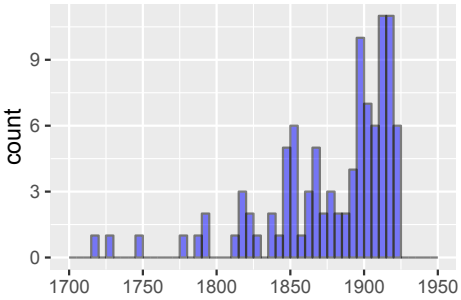
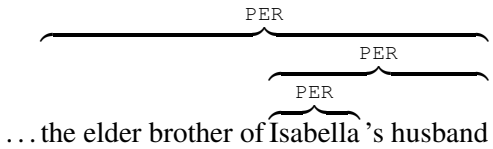


Figure 1: Distribution of texts in LitBank over time.

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:



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.

Category	n	Frequency
PER	24,180	83.1%
FAC	2,330	8.0%
LOC	1,289	4.4%
GPE	948	3.3%
VEH	207	0.7%
ORG	149	0.5%

Table 2: Counts of entity type.

Category	n	Frequency
PRON	15,816	54.3%
NOM	9,737	33.5%
PROP	3,550	12.2%

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)—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.

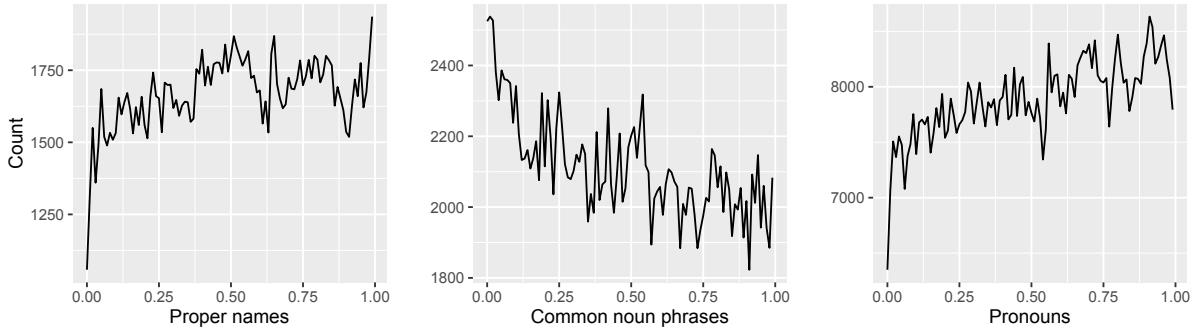


Figure 2: Distribution over narrative time of automatically predicted PROP, NOM and PRON person entities in 100 English-language books from Project Gutenberg, excluding all paratextual material from Project Gutenberg (legal boilerplate) and the print original (tables of contents, advertisements, etc.).

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.

Character	Text	n
Buck Mulligan	Ulysses	43
Convict	Great Expectation	33
Mrs. Bennett	Pride and Prejudice	29
Ragged Dick	Ragged Dick	28
Mr. Bennett	Pride and Prejudice	28

Table 4: Characters with most annotated dialogue.

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 diagetic 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-diagetic 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, hypothethicals, 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

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 consciouness (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

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. A comparison of the events and relations across ACE, ERE, TAC-KBP, and Framenet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-2907. URL http://aclweb.org/anthology/W14-2907.

Mark Algee-Hewitt. The machinery of suspense. http://markalgeehewitt.org/index.php/main-page/projects/the-machinery-of-suspense/, 2016.

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. Canon/archive: Large-scale dynamics in the literary field. Literary Lab Pamphlet 11, 2016.

Mieke Bal. *Narratology : Introduction to the Theory of Narrative*. University of Toronto Press, 2017.

David Bamman and Gregory Crane. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer, 2011.

David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1220. URL https://www.aclweb.org/anthology/N19-1220.

David Bamman, Olivia Lewke, and Anya Mansoor. An annotated dataset of coreference in English literature. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.6.

Yehoshua Bar-Hillel. The present status of automatic translation of languages. In *Advances in computers*, volume 1, pages 91–163. Elsevier, 1960.

Allan Bell. Language style as audience design. *Language in Society*, 13:145–204, 1984.

Katherine Bode. "man people woman life" / "creek sheep cattle horses": Influence, distinction, and literary traditions. In *A World of Fiction: Digital Collections and the Future of Literary History*, 2018.

Julian Brooke, Adam Hammond, and Graeme Hirst. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 26–35, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-2504.

Julian Brooke, Adam Hammond, and Timothy Baldwin. Bootstrapped text-level named entity recognition for literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–350, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2056. URL https://www.aclweb.org/anthology/P16-2056.

Bill Brown. Thing theory. *Critical Inquiry*, 28(1):1–22, 2001. ISSN 00931896, 15397858. URL http://www.jstor.org/stable/1344258.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. Deep learning for free indirect representation. *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 2019.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. Corpus REDEWIEDERGABE. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 803–812, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.100.

Alicia Burga, Joan Codina, Gabriella Ferraro, Horacio Saggion, and Leo Wanner. The challenge of syntactic dependency parsing adaptation for the patent domain. In *ESSLLI-13 Workshop on Extrinsic Parse Improvement*, 2013.

Nathanael Chambers. *Inducing Event Schemas and their Participants from Unlabeled Text*. PhD thesis, Stanford University, 2011.

Tanya Clement and Stephen McLaughlin. Measured applause: Toward a cultural analysis of audio collections. *Cultural Analytics*, 2018.

Tanya Clement, David Tcheng, Loretta Auvil, Boris Capitanu, and Megan Monroe. Sounding for meaning: Using theories of knowledge representation to analyze aural patterns in texts. *DHQ: Digital Humanities Quarterly*, 7(1), 2013.

Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Ken H Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

Leon Derczynski, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013a.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206, 2013b.

Caleb Derven, Aja Teehan, and John Keating. Mapping and unmapping Joyce: Geoparsing wandering rocks. In *Digital Humanities 2014*, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

David K. Elson and Kathleen R. McKeown. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1013–1019. AAAI Press, 2010.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 138–147, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Elizabeth F. Evans and Matthew Wilkens. Nation, ethnicity, and the geography of british fiction, 1880–1940. *Cultural Analytics*, 2018.

Mark A. Finlayson. ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory. *Digital Scholarship in the Humanities*, 32(2):284–300, 12 2015. ISSN 2055-7671. doi: 10.1093/llc/fqv067. URL `https://doi.org/10.1093/llc/fqv067`.

Mark Alan Finlayson. Inferring propp's functions from semantically annotated text. *The Journal of American Folklore*, 129(511):55–77, 2016. ISSN 00218715, 15351882. URL `https://www.jstor.org/stable/10.5406/jamerfolk.129.511.0055`.

Lucie Flekova and Iryna Gurevych. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1208`.

Charlotte Galves and Pablo Faria. Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/corpus/en/index.html, 2010.

Andrew Gelman and Thomas C. Little. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 1997.

Gérard Genette. *Figures of Literary Discourse*. Columbia University Press, New York, 1982.

Gérard Genette. *Narrative Discourse: An Essay in Method*. Cornell University Press, 1983.

Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202, 2001.

Dag TT Haug and Marius Jøhndal. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008), Marrakech, Morocco, 1st June 2008*, pages 27–34, 2008.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Natalie Houston. Enjambment and the poetic line: Towards a computational poetics. In *Digital Humanities 2014*, 2014.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Jennifer Isasi. *Posibilidades de la minería de datos digital para el análisis del personaje literario en la novela española: El caso de Galdós y los "Episodios Nacionales"*. PhD thesis, University of Nebraska, 2017.

Matthew Jockers. Revealing sentiment and plot arcs with the syuzhet package. `http://www.matthewjockers.net/2015/02/02/syuzhet/`, 2015.

Matthew Jockers and Gabi Kirilloff. Understanding gender and character agency in the 19th-century novel. *Cultural Analytics*, 2016.

Lauren Klein. Distant reading after moretti. `https://arcade.stanford.edu/blogs/distant-reading-after-moretti`, 2018.

Lauren F. Klein. Dimensions of Scale: Invisible Labor, Editorial Work, and the Future of Quantitative Literary Studies. *Pmla*, 135(1):23–39, 2020. ISSN 0030-8129. doi: 10.1632/pmla.2020.135.1.23.

Sheldon Klein, John F. Aeschlimann, David F. Balsiger, Steven L. Converse, Claudine Court, Mark Foster, Robin Lao, John D. Oakley, and Joel Smith and. Automatic novel writing. Technical report, University of Wisconsin-Madison, 1973.

Eve Kraicer and Andrew Piper. Social characters: The hierarchy of gender in contemporary English-language fiction. *Cultural Analytics*, 2018.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. Penn-Helsinki parsed corpus of Early Modern English. *Philadelphia: Department of Linguistics, University of Pennsylvania*, 2004.

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis. Description of a corpus of character references in German novels - DROC [Deutsches ROman Corpus]. DARIAH-DE Working Papers No. 27., 2017.

Matthew Lease and Eugene Charniak. Parsing biomedical literature. In *Natural Language Processing–IJCNLP 2005*, pages 58–69. Springer, 2005.

Hoyt Long and Richard Jean So. Turbulent Flow: A Computational Model of World Literature. *Modern Language Quarterly*, 77(3):345–367, 09 2016. ISSN 0026-7929. doi: 10.1215/00267929-3570656. URL `https://doi.org/10.1215/00267929-3570656`.

Hoyt Long, Anatoly Detwyler, and Yuancheng Zhu. Self-repetition and east asian literary modernity, 1900-1930. *Journal of Cultural Analytics*, 5 2018.

Marit J MacArthur, Georgia Zellou, and Lee M Miller. Beyond poet voice: sampling the (non-) performance styles of 100 american poets. *Cultural Analytics*, 2018.

Laura Mandell. Gender and Cultural Analytics: Finding or Making Stereotypes? In *Debates in Digital Humanities*, 2019.

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

David McClure. Distributions of words across narrative time in 27,266 novels. `https://litlab.stanford.edu/distributions-of-words-27k-novels/`, 2017.

Laura B. McGrath, Devin Higgins, and Arend Hintze. Measuring modernist novelty. In *Cultural Analytics*, 2018.

James R Meehan. TALE-SPIN, an interactive program that writes stories. In *IJCAI*, volume 77, pages 91–98, 1977.

T. C. Mendenhall. The characteristic curves of composition. *Science*, 1887.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

Taesun Moon and Jason Baldridge. Part-of-speech tagging for Middle English through alignment and projection of parallel diachronic texts. In *EMNLP-CoNLL*, pages 390–399, 2007.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. *NAACL*, 2016.

F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

Martin Mueller. Wordhoard. http://wordhoard.northwestern.edu/, Accessed 2015.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32, 07 2017a.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 460–470, 2017b.

Aakanksha Naik and Carolyn Rose. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7618–7624, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.681.

Janis Pagel and Nils Reiter. GerDraCor-coref: A coreference corpus for dramatic texts in German. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 55–64, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.7.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie plot analysis via turning point identification. 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)*, pages 1707–1717, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1180. URL https://www.aclweb.org/anthology/D19-1180.

Marco Passarotti. Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In Petrilli Raffaella and Femia Diego, editors, *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, Settembre 2006*, pages 187–205. Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio, 2007.

Viktor Pekar, Juntao Yu, Mohab El-karef, and Bernd Bohnet. Exploring options for fast domain adaptation of dependency parsers. *SPMRL-SANCL 2014*, page 54, 2014.

Marco Pennacchiotti and Fabio Massimo Zanzotto. Natural language processing across time: An empirical investigation on Italian. In *Advances in natural language processing*, pages 371–382. Springer, 2008.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Kenton Rambsy and Peace Ossom-Williamson. Lost in the city: An exploration of edward p. jones's short fiction. https://iopn.library.illinois.edu/scalar/lost-in-the-city-a-exploration-of-edward-p-joness-short-fiction-/index, 2019.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics (CL2007)*, 2007.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31, 2016.

Jonathan Reeve. The henry james sentence: New quantitative approaches. https://jonreeve.com/2017/06/henry-james-sentence/, 2017.

Nils Reiter, Marcus Willand, and Evelyn Gius. A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics*, 12 2019.

Roopika Risam. Other worlds, other DHs: Notes towards a DH accent. *Digital Scholarship in the Humanities*, 32(2):377–384, 02 2016. ISSN 2055-7671. doi: 10.1093/llc/fqv063. URL https://doi.org/10.1093/llc/fqv063.

Rachael G. Samberg and Cody Hennesy. Law and literacy in non-consumptive text mining: Guiding researchers through the landscape of computational text analysis. In *Copyright Conversations: Rights Literacy in a Digital World*, 2019.

Evan Sandhaus. The new york times annotated corpus. LDC, 2008.

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1970–1978, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.178.

Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett. Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 19–23. Association for Computational Linguistics, 2011.

Matthew Sims and David Bamman. Measuring information propagation in literary social networks, 2020.

Matthew Sims, Jong Ho Park, and David Bamman. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1353. URL https://www.aclweb.org/anthology/P19-1353.

R. Sprugnoli and S. Tonelli. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4):485–506, 2017. doi: 10.1017/S1351324916000292.

Beth M. Sundheim. Overview of the third message understanding conference. In *Processing of the Third Message Understanding Conference*, 1991.

Ann Taylor and Anthony S Kroch. The Penn-Helsinki Parsed Corpus of Middle English. *University of Pennsylvania*, 2000.

Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Parsed Corpus of Early English Correspondence. Oxford Text Archive, 2006.

Dennis Yi Tenen. Toward a computational archaeology of fictional space. *New Literary History*, 2018.

Ted Underwood. Why literary time is measured in minutes. Technical report, University of Illinois, 2016.

Ted Underwood. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019.

Ted Underwood, David Bamman, and Sabrina Lee. The transformation of gender in English-language fiction. *Cultural Analytics*, 2018.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. Are fictional voices distinguishable? classifying character voices in modern drama. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–34, Minneapolis, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2504. URL https://www.aclweb.org/anthology/W19-2504.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 multilingual training corpus. LDC, 2006.

Byron Wallace. Multiple narrative disentanglement: Unraveling infinite jest. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N12-1001.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. Ontonotes release 5.0. https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf, 2012.

David Wilmot and Frank Keller. Modelling suspense in short stories as uncertainty reduction over neural representation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1763–1788, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.161.

Erin Wolfe. Natural language processing in the humanities: A case study in automated metadata enhancement. *code4lib*, 46, 2019.