Multilingual BookNLP: Building a Literary NLP Pipeline Across Languages

Much work in the computational analysis of literature relies on pipelines in natural language processing to reason about the linguistic structure of text. BookNLP (Bamman et al., 2014)[1] is one such pipeline: when run on an input book, it tokenizes the text into sentences and words, and assigns each word a part-of-speech tag (such as noun or verb) and named entity category (person, location); it predicts the syntactic structure of each sentence (e.g., which words are the subject and direct object of verbs), performs pronominal coreference resolution (linking mentions of e.g. "she" or "he" to the characters they refer to), and identifies the set of unique characters from those mentions; it attributes quotes to their speakers, and then represents each character as the set of actions they do or have done to them, along with the objects they possess and the attributes that are predicated of them. While many existing tools such as Stanford CoreNLP (Manning et al., 2014) or Spacy[2] often struggle with books—where long, complex sentences strain the limits of syntactic parsers with super-linear computational complexity, and the sheer document length makes tasks like coreference resolution prohibitively expensive—BookNLP is natively designed to support the analysis of literature. Figure 1 illustrates this process for Dickens' *Great Expectations*, showing the layers of annotation for a single sentence from that text; here we can see the token *Pip* has been resolved to the character PHILIP PIRRIP, *Joe* to the character JOE GARGERY, and *she* and *her* to Pip's sister (known as *Mrs. Joe* throughout the text).

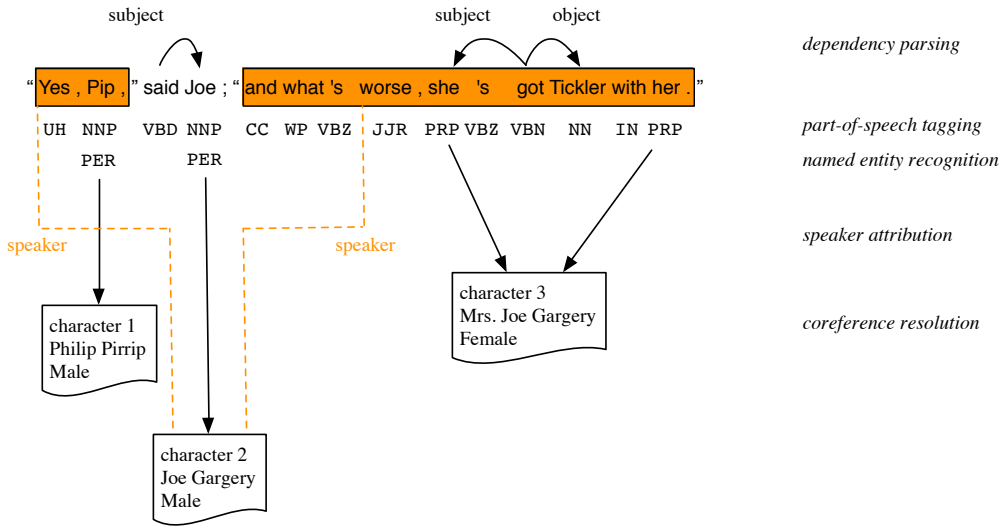

Figure 1: Sample BookNLP output illustrating different levels of annotation. For clarity, the only syntactic relations shown are subject and direct object.

The power of BookNLP comes in its role as an algorithmic measuring device; this device can be used to enable distant reading of texts, by aggregating information about the behavior of characters at a scale too large for a single individual to read, and to enable computer-augmented close reading, by highlighting the areas of idiosyncrasy and difference between characters in a single text. Table 1, for instance, illustrates the top actions that are most characteristic of Mr. Darcy and Elizabeth Bennet as agents in Austen's *Pride and*

[1] https://github.com/dbamman/book-nlp
[2] http://www.spacy.io

Prejudice, with respect to each other—here we can see a degree of focalization on Elizabeth that emphasizes her mental states (*cried*, *listened*, *found*, *think*, *felt*) in distinction to Darcy, who predominantly participates in verbs of action.

Mr. Darcy	Elizabeth
has	turned
say	cried
came	listened
called	help
doing	added
come	found
danced	think
have	felt

Table 1: The most characteristic actions associated with Mr. Darcy and Elizabeth Bennet from *Pride and Prejudice*, relative to each other.

The existence of this tool has driven much work in the computational humanities, especially surrounding character: Underwood et al. (2018) use it to measure the amount of attention given to characters as a function of the their gender (and that of the author) in 104,000 texts published over 170 years; Kraicer and Piper (2018) use it to explore the relative frequency of major and minor characters, along with the heteronormativity of their relationships, in 1,333 novels from the 21st century; Dubnicek et al. (2018) use it to analyze the representation of characters with disabilities; Ardanuy and Sporleder (2015) use it to analyze the relationship between character and literary genre; Wolfe (2019) uses it to characterize locations in novels written by African Americans in the Black Book Interactive Project; at the DH 2019 conference, Googasian and Heuser (2019) use it to model anthropomorphism in animal writing, while Cheng (2019) uses it to explore literary embodiment by analyzing the physical features of characters. BookNLP is widely used to operationalize texts in order to drive literary argument.

At the same time, however, BookNLP has one major limitation: it currently only supports texts written in English, further exacerbating what Roopika Risam notes is "the Anglophone focus of the field" of digital humanities (Risam, 2016). The goal of this project is to join others in shifting this narrow focus on English to include other languages as well, by developing versions of BookNLP for Spanish, Japanese, Russian and German, and creating a blueprint for others to develop it for further languages in the future.

We will structure our work under the scope of this grant in three stages: 1.) creating a minimum viable BookNLP for Spanish, Japanese, Russian and German from existing annotated resources (which are focused primarily on the domain of news); 2.) assessing the performance of BookNLP on literary data in each of those four languages and creating new annotated data specifically for the domain of literature when needed; and 3.) exploring the affordances of a specifically multilingual system; while each language-specific BookNLP will enable new research in the computational study of literature in that language, we will explore the kinds of research that are made possible in a specifically comparative setting, by running this system on a corpus of Spanish, Japanese, Russian, German and English texts to enable cross-linguistic cultural analytics.

1 Phase 1: Minimum Viable BookNLP

BookNLP is a trained system that relies on language-specific annotated data. In order to enable the core pipeline described above, this data must include texts that are annotated for part-of-speech, NER, syntax, and coreference. Fortunately, many of these layers of annotation already exist for each of these four languages; these datasets are generally focused on the domain of news, but have some variety in domain for different tasks (including fiction).

Spanish. For the main NLP tasks considered above, Spanish has 1 million words annotated for part-of-speech and dependency syntax, primarily drawn from newspapers, blogs and reviews (Taulé et al., 2008; McDonald et al., 2013), along with 500,000 tokens from newspapers annotated for NER and coreference (Taulé et al., 2008).

Japanese. Japanese has annotated data in the form of 1.1 million words from several genres in the Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2014)—including news, magazines, books, blogs, and textbooks—all annotated for part-of-speech and dependency syntax (Asahara et al., 2018), along with 127,000 tokens from news annotated for fine-grained NER (Mai et al., 2018) and 10,000 sentences from newspaper articles annotated for coreference, including zero anaphora (Kawahara et al., 2002). Japanese also requires word segmentation, for which we will explore existing tools such as MeCab,[3] Unidic[4] and RakutenMA,[5] along with more recent neural approaches such as Nagisa.[6]

Russian. Russian has several treebanks totaling 1.1 million words from a variety of genres, including fiction, news, and academic articles, all annotated for part-of-speech and dependency syntax (Droganova et al., 2018), along with 44,000 tokens of news annotated with NER (Gareev et al., 2013) and 156,000 tokens predominantly of news, essays and fiction annotated for coreference (Toldova et al., 2014).

German. German has the most resources of the four target languages, with 3.8 million words of news (Borges Völker et al., 2019; Foth et al., 2014) annotated for part-of-speech and dependency syntax and 590,000 tokens from news and Wikipedia annotated for NER (Benikova et al., 2014). German is also unique in having annotations for literary texts, including 393,000 tokens from 90 German novels annotated for character coreference (Krug et al., 2017) and 489,459 tokens annotated for speech, thought and writing, including direct, indirect, free indirect and reported speech.[7]

Phase one of this work will focus on leveraging these existing resources to build a functioning BookNLP system for each of these languages within a common infrastructure. At a minimum, this includes building statistical models for the core problems of part-of-speech tagging, dependency parsing, named entity recognition and coreference resolution trained on these existing resources, and developing algorithms for the deterministic components of speaker attribution, character clustering, and character gender inference. While the learned tasks of POS tagging, parsing, NER, and coreference resolution can be natively trained on in-language data (and can learn, as a result of that training, how coreference in English behaves differently from coreference in Russian), part of the complexity of this stage will be in adapting the deterministic components to the nuances of each individual language, given language-specific differences in both dialogue and characterization. A deliverable at this stage will be functional BookNLP systems for our four target languages.

2 Phase 2: Optimizing BookNLP for literature

The majority of datasets outlined above contain annotations for text drawn from news—either newswire, newspaper articles, or online news sources. Table 2 provides a summary of recent research that has investigated the disparity between training data and test data for several NLP tasks (including part-of-speech tagging, syntactic parsing, named entity recognition and coreference resolution). While many of these tools are trained on the same fixed corpora (comprised primarily of newswire), they suffer a dramatic drop in performance when used to analyze texts that come from a substantially different domain. Without any form of adaptation (such as normalizing spelling across time spans), the performance of an out-of-the-box part-of-speech tagger can, at worse, be half that of its performance on contemporary newswire. On average,

[3] https://taku910.github.io/mecab/
[4] https://unidic.ninjal.ac.jp/back_number#unidic_cwj
[5] https://github.com/quinnanya/japanese-segmenter
[6] https://github.com/taishi-i/nagisa
[7] https://github.com/redewiedergabe/corpus

differences in style amount to a drop in performance of approximately 10-20 absolute percentage points across tasks. These are substantial losses, and can have the effect of 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. (2019b)	Nested NER	News	68.8 F	English literature	45.7 F
Bamman et al. (2019a)	Coreference	News	83.2 F	English literature	72.9 F

Table 2: Out-of-domain performance for several NLP tasks, including POS tagging, phrase structure (PS) parsing, dependency parsing, named entity recognition and coreference resolution. Accuracies are reported in percentages; phrase structure parsing, NER and coreference are reported in F1 measure.

Much work in the NLP community has focused on narrowing this gap in performance between training data and test data, including domain adaptation (Blitzer et al., 2006; Yang, 2017), active learning (Settles, 2012) and bootstrapping low-resource languages with fixed annotation budgets (Garrette and Baldridge, 2013), including work focused explicitly on domain adaption for historical texts (Wing, 2015; Yang and Eisenstein, 2016). However, one of the most proven methods for increasing performance in a given domain is to simply annotate more data within that domain—a process that eliminates the bottleneck of requiring computational expertise and puts control for performance in the hands of domain experts. Where training data is available within the target domain, it can substantially increase performance, almost to levels approaching state-of-the-art on English newswire. In-domain annotations have increased POS tagging accuracy on Early Modern German texts from 69.6% to 91.0% (Scheible et al., 2011) and on Middle English texts from 56.2% to 93.7% (Moon and Baldridge, 2007). The PI's work has demonstrated this phenomenon for specifically literary texts as well; the creation of a new dataset of annotations for 100 works of English fiction increased performance on the task of NER from an F-score of 45.7 to 68.3 (Bamman et al., 2019b), and for coreference resolution from 72.9 to 79.3 (Bamman et al., 2019a).

While phase one will deliver functional BookNLP systems for Spanish, Japanese, Russian and German, phase two will carry out an assessment of their performance on literary texts in particular, evaluating the individual components of the pipeline to measure the drop in performance we expect will take place when training on news and testing on literature. This assessment will illuminate the specific areas that require in-domain annotations to improve performance. To execute this phase, we will annotate a small sample of texts for each subtask (part-of-speech tagging, parsing, NER, coreference resolution, speaker attribution, character name clustering and character gender inference) for each of the four languages, carried out by undergraduate and graduate researchers at UC Berkeley, in consultation with our advisory board; we will design our annotation strategy at this stage with a power analysis to only annotate as much data as is required to get a reasonable estimate of performance on literary texts. Once we assess which subtasks for which languages are most in need of improvement, we will carry out annotations specifically targeting those. Given the size and range of existing resources, likely candidates for this may include named entity tagging

and coreference resolution in Russian and Japanese, which have the smallest resources compared to other languages; additionally, while Spanish, Russian and Japanese all elide pronouns as subjects of verbs when they are pragmatically inferable, Japanese also elides them in other grammatical positions (such as direct objects), making coreference especially challenging. Focusing on data at this stage will allow us to build on the functional system developed in phase one by simply adding more targeted data to train on. Deliverables at this stage include newly annotated data, BookNLP systems that improve their performance relative to phase one, and a blueprint for other researchers to build versions of BookNLP for new languages beyond those considered in the scope of this project.

3 Phase 3: Multilingual affordances

The work carried out in phases one and two essentially treat each language in isolation, building BookNLP systems for Spanish, Japanese, Russian and German that consider performance and identify data to annotate within each specific language. Phase three will explore the unique affordances that placing all languages within a common infrastructure makes possible. We will explore this in two dimensions: technically, implementing a new BookNLP feature available for one language (such as scene segmentation in German) to the other three languages; and running our improved models on a corpus of Spanish, Japanese, Russian, German and English texts to enable cross-linguistic cultural analytics. We will consult with our advisory board on suitable topics for analysis at this stage, but possibilities include:

- A cross-linguistic comparison replicating previous work measuring attention as a function of character gender (Underwood et al., 2018; Kraicer and Piper, 2018) to analyze disparities in representation across different linguistic traditions. Do literatures in Spanish, German, Japanese and Russian display a similar disparity in the representation of men and women characters, or are there other dynamics at play? By analyzing the words associated with specific genders (i.e., the verbs most common to female characters), can we understand how gender is marked differently in different literary traditions?

- An analysis of characterization with respect to questions of interiority and focalization (Long et al., 2018; Piper, 2018; Jannidis, 2004; Eder et al., 2010). How do different cultures convey the psychological investments of literary characters, whether in terms of the verb types associated with their actions (cognitive, stative, dynamic) or the embedding of characters within other characters' points of view (i.e. the networks of focalization created when characters are observed or discussed by others)?

- An analysis of familial structures across different literatures. While BookNLP allows us to identify the characters present in a narrative, there is much work in attempting to infer the relationships between them—including not only whether a relationship is positive or negative (Chaturvedi et al., 2017; Krishnan and Eisenstein, 2015) but also the specific familial structure, such as MOTHER–DAUGHTER, HUSBAND–WIFE or BROTHER–BROTHER (Makazhanov et al., 2014). How are families represented in literary texts, and how do those representations differ across different linguistic traditions?

For the analysis selected, we will leverage the large-scale collection of digitized books in the HathiTrust Research Center, which currently includes 243,565 texts in Japanese, 781,306 texts in German, 597,609 works in Spanish, and 304,180 volumes in Russian, in addition to 4.6 million volumes in English.[8] These texts have all been OCR'd from printed books, with variable OCR quality (especially for the languages of Japanese and Russian). To ensure high-quality data for those two languages in particular, we will supplement the HathiTrust data with two additional datasets: first, the Aozora Bunko collection,[9] which contains

[8] https://www.hathitrust.org/visualizations_languages

[9] https://www.aozora.gr.jp

approximately 15,000 literary and non-fiction works in Japanese that are hand-coded and checked for consistency; and second, the Maksim Moshkow library of Russian texts,[10] which contains a variety of works of Russian literature, both modern and traditional. Both resources can be seen as the equivalent to Project Gutenberg—both in their community engagement and in the scope of their materials in the public domain.

We will identify works of fiction for each language in this collection using the methodology of Underwood (2014) and further manually narrow each language collection to comparable sets (e.g., equivalent time periods or genres), depending on the specific research question being examined. We will then pass each text through our multilingual BookNLP pipeline to enable comparison across different literary traditions.

Our goal in this stage will be to demonstrate the value of comparative computational analysis afforded by this work, and will target publication in a literary venue (such as the *Journal of Cultural Analytics* or *Critical Inquiry*). The application of BookNLP across multiple languages will initiate a new research program aimed at understanding cross-cultural differences and similarities around the construction of literary character. While we have a great deal of qualitative literature to this effect from the field of comparative literary studies, there are as yet no empirical and quantitatively driven studies that address the consistency or variability of the practice of characterization across different literary cultures. Knowledge of how homogenous or divergent the construction of character, familial relations, or psychological profiles are across geographically distant cultures can give us insights into the nature of human narrative and its cultural variability.

4 Enhancing the Humanities

As mentioned above, BookNLP in English has supported a range of emerging research in the digital humanities. The most immediate impact of expanding BookNLP to Spanish, Japanese, German and Russian is to enable the kind of empirical studies already undertaken in English to literatures in those languages as well—where any of the studies described above could be replicated on (e.g.) Spanish literature to test if similar findings hold. This has the effect of broadening the set of computational tools available for the study of non-English texts.

While much work has pointed out the Anglophone-centered nature of the digital humanities communities, this critique is often supported by analyses of publication language (Fiormonte, 2015), the distribution of institutional DH centers (Terras, 2011), and the geographic distribution of authors at major DH conferences (Weingart, 2016); but a similar concentration of energy exists for language-specific tools as well: while the popular Stanford CoreNLP offers a full pipeline for English NLP, it only offers a subset of functionality (part-of-speech tagging and parsing) for other languages, omitting critical components such as coreference resolution. And despite the Anglophone focus and use of English as a lingua franca in DH (Gil and Ortega, 2016), there is a wealth of research in the digital humanities applying computational methods to non-English literatures—fostered by local organizations such Digital Humanities im deutschsprachigen Raum, Humanidades Digitales Hispánicas, the Japanese Association for Digital Humanities, the European Association for Digital Humanities and Red de Humanidades Digitales, among others, and cross-cutting interest groups like GO::DH—that could benefit from the development of literary NLP pipelines specifically designed to support their language.

More broadly, however, year three of this project is focused on the affordances of adopting a specifically *multilingual* point of view in computational approaches to literary analysis—that is, exploring the possibility for literary insight when reasoning about multiple literatures together. In doing so, this project will contribute to the important movement of world literature studies (Damrosch, 2008, 2014), helping break down national barriers in our understanding of human storytelling. By putting each language tradition through the same pipeline, and carrying out the same measurements on each text regardless of language, we can enable comparative analysis across different literary, cultural and linguistic traditions in a digital

[10] `http://lib.ru/`

context. We see this effort as setting in motion the expansion to further languages, enabling the increasingly multilingual and collaborative study of literature.

5 Environmental Scan

The broad work envisioned under the scope of this grant involves assembling a natural language processing pipeline in four languages specifically optimized for literary texts; this draws on related work in NLP pipelines generally and previous work exploring the value of in-domain annotated data for literature.

NLP pipelines. BookNLP is a natural language processing pipeline developed to support the analysis of literary texts in particular. Several other NLP pipelines exist that could potentially be adapted for literary texts. Stanford CoreNLP (Manning et al., 2014) and Spacy[11] are both established pipelines for part-of-speech tagging, parsing, named entity recognition, coreference resolution, and speaker attribution, but have two drawbacks: they are both trained on news texts (including the benchmark OntoNotes and CoNLL datasets) and optimized for short news-length articles as well; the computational complexity of phrase-structure parsing in CoreNLP is cubic in the length of the sentence, and is simply unable to parse documents with long sentences (such as the work of Henry James); the complexity of coreference resolution likewise is quadratic in the number of entities in a text, which grinds to a halt with novel-length texts. NLTK is a similarly widespread library for a variety of NLP tasks (including tokenization, named entity recognition, and an interface with WordNet), but does not offer trained models for other important elements in the pipeline (such as parsing and coreference resolution). GutenTag (Brooke et al., 2015) is a tool specifically designed for processing texts from Project Gutenberg; while it provides a framework for including tagging options such as part-of-speech tagging and quotation attribution, one of its most useful methods in the context of fiction is the ability to reason about chapter boundaries; we will explore this tokenization choice in our work as well. Finally, as more and more elements of natural language processing demonstrate improvements in performance as a result of using contextual word representations such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), the HuggingFace Transformer library (Wolf et al., 2019) is another important resource that we will incorporate into our own pipeline; this library currently offers trained models for English and a general multilingual model, along with user-supplied models optimized for specific languages such as Japanese and German.

In-domain annotations for literature. BookNLP relies on domain-specific annotated texts to improve performance; while much of the effort in this project comes in creating annotated data for both evaluation and training, a variety of annotated corpora have been created to support NLP tasks for fiction in different languages; these include the following:

- The DROC corpus, which contains coreference annotations for selections from 90 German novels published between the 17th and 20th centuries (Krug et al., 2017). Additionally, 489,459 tokens have been annotated for speech, thought and writing, including direct, indirect, free indirect and reported speech.
- The Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Faria, 2010), which contains part-of-speech and syntactic annotations for 2.8 million words from texts dating from 1380-1881.
- WordHoard, which contains part-of-speech annotations (automatically assigned and manually revised) for Shakespeare, Chaucer and Spenser (Mueller, 2015)

[11] `http://www.spacy.io`

- The Perseus Greek and Latin treebanks (Bamman and Crane, 2011), which contain morphosyntactic annotations for classical Greek and Latin works; the Index Thomisticus (Passarotti, 2007), which contains morphosyntactic annotations for the works of Thomas Aquinas; and the PROIEL treebank (Haug and Jøhndal, 2008), which contains similar annotations for several translations of the Bible (Greek, Latin, Gothic, Armenian and Church Slavonic).

- Several parsed corpora of historical English, which include morphosyntactic annotations of texts from Old English to the Early Modern era. These include The Penn-Helsinki Parsed Corpus of Middle English (Taylor and Kroch, 2000), The Parsed Corpus of Early Modern English (Kroch et al., 2004), the York-Toronto-Helsinki Parsed Corpora of Old English Prose and Old English poetry and the Parsed Corpus of Early English Correspondence (Taylor et al., 2006).

- The Icelandic Parsed Historical Corpus (Rögnvaldsson et al., 2012), which contains annotations of texts dating from 1100–the present.

Where possible, the open-access datasets among this set will be incorporated into the systems designed here (such as the DROC corpus for German); the availability of annotated data for other languages (such as Latin and Greek) also paves the way for the future development of BookNLP systems for those languages as well.

6 History of the Project

The work proposed here draws on a history of research along two dimensions. First, BookNLP was initially developed and published by the PI in 2014 (Bamman et al., 2014) and has been used by a range of work over the past six years, both within the digital humanities (Underwood et al., 2018; Kraicer and Piper, 2018; Dubnicek et al., 2018; Ardanuy and Sporleder, 2015; Wolfe, 2019; Googasian and Heuser, 2019; Cheng, 2019) and in natural language processing more broadly (Iyyer et al., 2016; Muzny et al., 2017; Chaturvedi et al., 2018; Zhang et al., 2019). For this entire time, BookNLP has been available exclusively for English, leading to advances in the computational analysis of literature within this language, but excluding that for literatures in other languages.

Second, this work draws on the PI's related research in building LitBank, a dataset to improve natural language processing specifically within the domain of literature. This has resulted in the creation of datasets for entity recognition (Bamman et al., 2019b), event detection (Sims et al., 2019) and coreference resolution (Bamman et al., 2019a) that each dramatically improve the state-of-the-art for components of the NLP pipeline compared to models trained and optimized for news.

7 Work plan

The specific details of the work plan align with the phases outlined above; each phase will take one year to complete.

7.1 Year 1

In year one, we will focus on building functional BookNLP systems for our four target languages of Spanish, Japanese, German and Russian. This will be carried out by a graduate student researcher at UC Berkeley, who will identify available resources in each of those target languages (in consultation with the PI and the advisory board for this project) and build a common Multilingual BookNLP architecture that can be used as the customizable foundation for all languages. This architecture will include a Python implementation

of trainable components for the tasks of part-of-speech tagging, syntactic parsing, named entity recognition and coreference resolution that make use of recent advancements in natural language processing, including the use of pre-trained contextual word embeddings like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) and contain implementations of algorithms for sequence labeling (used for part-of-speech tagging and named entity recognition), neural dependency parsing, and coreference resolution (Lee et al., 2017). These algorithms can be used for any language and can be trained simply given language-specific training data. In addition to these trainable components, this architecture will also contain hooks for the deterministic language-specific components of tokenization, quotation attribution, character clustering and character gender inference. A **deliverable** at this stage will be open-source BookNLP systems, published on Github, for each of those four languages.

7.2 Year 2

In year two, we will focus on improving the Multilingual BookNLP systems for the specific domain of literature. This will involve creating new annotated test data for each of the tasks outlined above, which will be carried out by undergraduate students at UC Berkeley with language expertise in Spanish, Japanese, German or Russian, supervised by a graduate researcher at UC Berkeley and the PI, in consultation with our advisory board.

August 2021–April 2022. For each language, we will annotate approximately 10,000 tokens, which for the task of POS tagging will allow us to measure accuracy with a 95% confidence interval within $\pm 0.5\%$. For each language, this will result in 50,000 annotated tokens (10,000 each for the five different token-level tasks of POS tagging, parsing, NER, pronominal coreference resolution and quotation attribution); and character-level annotations for 1,000 characters (for character name clustering and gender identification). This will result in an annotated test dataset of 200,000 tokens and 4,000 characters across all languages in total. For reference, our prior work annotating a literary dataset of 210,532 tokens for a single task took two undergraduates 3 months working at 10 hours per week. In our case here, we budget 3 times as long (9 months) for the same magnitude of data in order to accommodate the context shift in annotating several different tasks at once.

May 2022–July 2022. The evaluation part of phase 2 will allow us to determine which tasks for which languages are most in need of improvement for the BookNLP pipeline. With this information, we will select two of the worst-performing task/language pairs to improve by annotating more data, and create new datasets measuring approximately 200,000 tokens that we can use to train better systems. Given the modular, trainable BookNLP systems developed in year 1, we will train BookNLP on this new data (led by a graduate researcher at UC Berkeley).

Deliverables at this stage will include new datasets—test data for all tasks for all four languages, and training data for two tasks in two languages—published on Github under a Creative Commons Attribution 4.0 International License, and improved BookNLP models for two languages.

7.3 Year 3

In year three, we will have improved BookNLP systems for five languages (the four target languages here, along with English). We will focus in this year on two tasks: a.) documenting the processes undertaken during years 1 and 2 to enable others to build and train BookNLP systems for new languages beyond those studied here; and b.) exploring the affordances of reasoning about character using BookNLP across several languages. We will work in consultation with our advisory board to determine the best research questions

to investigate given the performance of the BookNLP systems developed during the first two years. Possibilities include replicating previous work in English on literary traditions of Spanish, Japanese, German and Russian, a cross-cultural analysis of interiority and focalization, and the representation of families in literary texts. All work will be undertaken by a graduate research assistant and undergraduate researchers at UC Berkeley, advised by the PI in consultation with our advisory board.

Deliverables at this stage will be publications in academic conferences (such as *Digital Humanities*), journals (such as *Cultural Analytics*, *Critical Inquiry*, and *Digital Humanities Quarterly*), and whitepapers.

8 Staff

David Bamman (UC Berkeley) will direct this project, directly supervising students at UC Berkeley. One graduate student at UC Berkeley will be funded for a continuous period of three years; their primary responsibility will be in advancing the technical work described above. This work will also fund undergraduate student annotators at UC Berkeley, who will be responsible for creating new annotated datasets in each of the target languages.

In order to build a functioning BookNLP system for the diverse languages of Spanish, Japanese, Russian and German, we will draw on the expertise of our advisory board, who bring deep knowledge of the linguistic properties of the language, knowledge of the literary tradition in the language, and deep experience using computational methods to drive literary inquiry. The primary roles of the advisory board will include a.) advising on linguistic properties of their language of expertise in the development of individual NLP components; b.) advising on annotation guidelines for specific linguistic phenomena; c.) advising on corpus selection for texts to annotate; d.) potentially connecting us with other language speakers to test or annotate the systems that are developed; e.) advising on existing data, resources, and methods in the language of their expertise that may be useful to incorporate; and f.) advising on research questions that can be addressed with this pipeline, either for their specific language of expertise, or in a cross-linguistic comparison.

Our advisory board will be comprised of two advisors for each target language: for **Spanish**, Jennifer Isasi (UT-Austin/Penn State) and Lucia Donatelli (Saarland University); for **German**, Andrew Piper (McGill University) and Fotis Jannidis (Universität Würzburg); for **Japanese**, Hoyt Long (University of Chicago) and Miyako Inoue (Stanford University); and for **Russian**, Quinn Dombrowski (Stanford University) and Andrew Janco (Haverford College).

9 Final Product and Dissemination

There will be three categories of work products originating in this grant. First, this work will directly result in open-source software for versions of BookNLP for the languages of Spanish, Japanese, Russian and German, along with documentation for how to use it; this software will be made publicly available on Github for others to freely use. Second, this work will result in the publication of data for a variety of NLP subtasks (including part-of-speech tagging, dependency parsing, named entity recognition, coreference resolution, quotation attribution, character name clustering and character gender inference) for our four target languages; this data will serve as a benchmark for others working on literary NLP in this space, and will also be made publicly available under a Creative Commons Attribution-4.0 license and published on Github. Finally, this work will result in publications in academic conferences (such as *Digital Humanities*, ACL, and EMNLP), journals (such as *Cultural Analytics*, *Critical Inquiry*, *Digital Humanities Quarterly*, and *TACL*), and whitepapers. These publications will document our methodology and present any empirical results in sufficient detail as to allow for replication.