Gutenberg+: A More Temporally Faithful Corpus for Diachronic NLP

Leon Hammerla, Alexander Mehler
Goethe University Frankfurt, Text Technology Lab
Robert-Mayer-Str. 10, 60325 Frankfurt
{hammerla, mehler}@em.uni-frankfurt.de

Abstract

We introduce Gutenberg+, a temporally more faithful version of the Project Gutenberg (PG) corpus, one of the most widely used resources for diachronic text analysis. Despite its popularity, the PG corpus contains a major yet overlooked flaw: around 15% of its entries are collections (e.g., anthologies of books, letters, or poems) rather than atomic works, which distorts temporal analyses since such collections may span multiple decades. We present an automatic method to detect and split these collections into their constituent works, producing a finer-grained and temporally consistent corpus. We further re-annotate publication years using LLM-based retrieval-augmented generative methods, demonstrating the potential of LLMs to enhance structured linguistic resources. To illustrate the utility of Gutenberg+, we conduct a small-scale diachronic case study on negation, showing that our refined corpus captures more nuanced cross-linguistic variation than the original PG data. Finally, we release the corpus in UIMA format with full metadata and linguistic annotations, providing a standardized resource for future research on diachronic language change.

Keywords: Corpus, Digital Libraries, Multilinguality, Text Analytics

1. Introduction

In the current LLM-dominated research landscape, large-scale unstructured corpora are widely used for training and evaluation. However, the reliability of downstream linguistic analyses, such as diachronic text analysis (DTA), critically depends on the structural and metadata integrity of these resources. While scale has increased dramatically, inconsistencies in document structure and coarse-grained temporal metadata can distort empirical findings and compromise reproducibility. Among such resources, the Project Gutenberg (1971-2026) (PG) collection stands out for its scope and accessibility, encompassing thousands of digitized literary works that, while technically spanning more than a millennium of writing, provide substantial and continuous coverage for roughly the past four centuries of language use (Gerlach and Font-Clos, 2020). Although the PG corpus holds great potential, its utility for DTA is constrained by noisy metadata, particularly incomplete or inaccurate publication information. These limitations hinder reliable temporal analysis and prevent the corpus from fully realizing its potential as a historical resource. Recent efforts have sought to address this problem and improve the temporal reliability of the PG corpus. For instance, Hegde et al. (2025) introduced CHRONOBERG, a temporally structured corpus of English book texts spanning 250 years, curated from Project Gutenberg and enriched with external temporal metadata from sources such as OpenLibrary and Wikipedia. Similarly, Momen et al. (2025) tackled the challenge of missing temporal annotations by leveraging LLMs with retrieval-augmented generation to estimate the production years of all PG books, potentially enhancing the corpus's usability for diachronic research. While these endeavors represent significant progress, they share a critical limitation: out of 76,534 items in the PG corpus, 11,650 ($\sim 15\%$) are collections, such as volumes of poems, anthologies, collected works, or books of letters. These collections often span multiple decades, meaning that their temporal annotations do not accurately reflect the underlying texts. To properly study the PG corpus through a diachronic lens, such collections must be split into their constituent items, creating a more fine-grained and temporally faithful resource. We address this structural limitation by transforming PG into a finer-grained, temporally consistent, and richly annotated resource suitable for reliable linguistic analysis and evaluation. As a test case, we quantify the impact of our more fine-grained corpus view through the lens of negation, a core linguistic phenomenon characterized by its dynamic nature, making it particularly suitable for examining language change over time (e.g. Jespersen et al., 2025; Croft, 1991; Mazzon, 2016). Our contributions are threefold:

1. We split the collections in the PG corpus into atomic parts, creating a finer-grained, temporally faithful representation of the texts.

2. We re-annotate the new split using the approach of Momen et al. (2025) and add negation annotations across nine languages using the D-Neg tool (Hammerla et al., 2025), which extends the NegBERT framework (Khandelwal and Sawant, 2020). We make this re-

Proceedings of the Workshop on Structured Linguistic Data and Evaluation, pages 86–92
May 11, 2026.

source publicly available for the research community.

3. We demonstrate the analytical impact of our structural refinement through a multilingual diachronic case study on negation, showing that fine-grained metadata significantly enhances the prediction of temporal distributions.

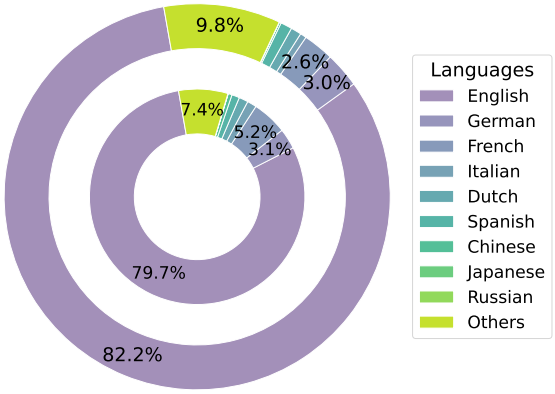

Figure 1: Distribution of items by language in the PG corpus, showing all languages for which negation can be detected, with the remaining languages grouped into the "Others" category. The inner ring illustrates the initial PG split; the outer ring depicts the fine-grained split produced by our version of the PG.

2. Dataset

2.1. Project Gutenberg

Our base corpus is derived from Project Gutenberg (1971-2026), a large-scale digital library of public-domain texts digitized by volunteers. The corpus comprises 76,534 items in 70 languages. English (61,016), French (4,010), Finnish (3,414), German (2,357), and Italian (1,075) are the most frequent languages, together accounting for about 94% of all items. The corpus is further organized into 474 virtual bookshelves. The largest categories are Novels (23,676 items), British Literature (9,758), Adventure (8,238), American Literature (7,584), and Children & Young Adult Reading (6,713). In addition, Project Gutenberg provides subject metadata, containing 41,423 distinct subject annotations. Among the most frequent are the Library of Congress Classification (LCC) labels[1] PS (12,068), PR (10,705), PZ (7,937), PQ (5,327), and PT (3,295).

2.2. Detecting and splitting Collections

To identify and split collections within the PG corpus, we first detect which items constitute collections. We leverage the LLM-generated summaries available in the corpus metadata for this purpose. Manual inspection revealed that PG items whose summaries contained the word collection within the first three sentences were almost always collections. A small-scale human evaluation (100 samples) confirmed this heuristic with 98% precision, which we prioritize to ensure that predicted positives are highly reliable. Any such text collection hinders the application of procedures that operate item-wise, such as publication year annotators. Therefore, mapping the members of these collections to individual items is indispensable.

After identifying the collections, we proceed in two steps. (1) We extract the table of contents of each collection using Algorithm 1, obtaining the titles of the individual items contained within. (2) We then split the collection text into its atomic parts using Algorithm 2. This process involves normalizing the text and detecting lines that directly match an item title surrounded by newline characters, which typically indicate a section heading. To assess the reliability of our collection-splitting procedure, we conducted a small-scale human evaluation on 100 randomly selected collections and achieved an accuracy of 86% for correctly segmented (i.e., satisfying) splits. The two primary sources of error were: (1) collections lacking a clear table of contents, which prevented the identification of distinct items (57% of errors), and (2) cases where item titles were detected from the table of contents but could not be matched to the corresponding sections in the text due to missing or inconsistent headlines (29% of errors). Given the absence of a standardized digitization format in the PG corpus, structural inconsistencies across editions remain an unavoidable limitation.

2.3. Publication Year Annotation

We refine the temporal dimension of the complete PG dataset by reusing the LLM-based retrieval-augmented publication year annotation method proposed by Momen et al. (2025), which has proven to be both effective and efficient. Specifically, we adopt their setup using gemma-2-9b-it with 8-bit quantization and the same prompting configuration. Following their best-performing setting, we provide only the first and last page of each item as context for year prediction. Since most PG items lack explicit page boundaries, we approximate page length by taking the first and last 3,000

[1]The reported LCC labels denote broad literature classes: PS = American literature, PR = English literature, PZ = fiction and juvenile belles lettres, PQ = French, Italian, Spanish, and Portuguese literature, and PT = German, Dutch, and Scandinavian literature.

Input: Collection text T, title t, search window w
Output: Chapters C, remaining text R
1. Split T into lines L; restrict search to first w lines.
2. Find first short line in L containing keyword from KEYWORDS $\rightarrow$ toc_start.
3. If not found and t contains a comma, return $[(t)], T$.
4. From toc_start onward, collect non-empty lines l with $2 < |l| < 40$, remove trailing numbers.
5. Stop after three consecutive empty lines.
6. Remove entries in PAGE_WORDS.
7. Split C into main chapters (ending with ":") and subchapters.
8. Return whichever set is non-empty along with remaining text.

Algorithm 1: Extract Table of Contents

Input: Collection text T, chapter titles C
Output: Items I, filtered chapters C', line ranges R
1. Split T into lines L.
2. For each chapter c_i in C, find candidate start lines using FINDNORMALIZEDLINE.
3. Keep the first valid start appearing after previous ones; record indices of invalid matches.
4. Append end of L to list of start positions.
5. Segment L between successive start positions $\rightarrow$ items I.
6. Remove unmatched chapters; assert $|I| = |C'|$.
7. Return I, C', and line ranges R.

Algorithm 2: Split Collections into Items

characters of each text.

2.4. Segmentation and Negation Annotation

For sentence and word segmentation, we use spaCy (Honnibal et al., 2020), employing its small model family[2]. For the negation detection, we use D-Neg (Hammerla et al., 2025). D-Neg expands the NegBERT architecture (Khandelwal and Sawant, 2020) for multiple languages and refines the classification using syntactic features such as part-of-speech tags and dependency trees, which we also extract with spaCy. The multilingual transformer backbone of D-Neg is EuroBERT (Boizard et al., 2025). The model is implemented as a token-classification pipeline. In a first pass, it identifies negation cues in the input sequence. In a second pass, for each detected cue, it predicts the scope of that cue by classifying each token in the sentence with respect to whether it falls within the corresponding negation scope. This procedure allows us to detect both negation cues and their associated scopes for 9 languages (see Figure 1). The detected cue types include affixal negation, negative adverbs, determiners, pronouns, prepositions, verbal negation cues, and discontinuous cues.

2.5. Corpus Statistics and Release

While the original PG corpus contained 76,534 entries, splitting its 11,650 collections (15%) expanded our Gutenberg+ corpus to 215,960 individual items. After segmentation, our corpus comprises 217.15 million sentences and 5.35 billion tokens, spanning one millennium of writing (see Figure 2). We identify 51.51 million instances of negation across nine languages, covering over 90% of all items in both the original PG corpus and our fine-grained version (see Figure 1). For distribution of our refined corpus we chose UIMA because it provides a standardized framework for representing and exchanging richly annotated text corpora (Ferrucci and Lally, 2004). UIMA facilitates integration with a variety of NLP tools and workflows, making it straightforward for other researchers to load, process, and extend our annotated corpus [3].

3. Experiments and Analysis

We propose two complementary approaches to evaluate the impact of our fine-grained split: comparing item frequency over time between the original and refined corpora, and examining language-specific differences in negation distributions. For all pairwise comparisons between distributions P and Q, we employ the Jensen–Shannon Divergence (JSD), a symmetric and bounded information-theoretic measure that generalizes the Kullback-Leibler divergence and is applicable to both univariate and multivariate, discrete, continuous, and mixed distributions:

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M),$$

where $M = \frac{1}{2}(P + Q)$, and KL denotes the Kullback-Leibler divergence. To correct for bias in the JSD estimate, we use a bootstrap-based bias correction (Efron and Tibshirani, 1994). Let $\mathrm{JSD}(P \parallel Q)$ be the original JSD estimate from the

[2] `en_core_web_sm`, `de_core_news_sm`, etc., depending on the language. If no language-specific model is available, we fall back to the English model.

[3] Our dataset is publicly available on HuggingFace and can be downloaded from `https://huggingface.co/datasets/PG-S/PG_FG`.

sample data, and let $\overline{\mathrm{JSD}}_B$ be the mean of the bootstrap JSD estimates. The bias is then estimated as:

$$\text{Bias} = \overline{\mathrm{JSD}}_B - \mathrm{JSD}(P \parallel Q).$$

The bias-corrected JSD estimate is:

$$\mathrm{JSD}_{\text{corrected}}(P \parallel Q) = \mathrm{JSD}(P \parallel Q) - \text{Bias}$$

$$= 2 \cdot \mathrm{JSD}(P \parallel Q) - \overline{\mathrm{JSD}}_B.$$

The bias-corrected $100(1-\alpha)\%$ confidence intervals for the JSD are computed as:

$$\mathrm{CI}_{\text{lower, corrected}} = 2 \cdot \mathrm{JSD}(P \parallel Q) - \mathrm{JSD}_{B,1-\alpha/2},$$

$$\mathrm{CI}_{\text{upper, corrected}} = 2 \cdot \mathrm{JSD}(P \parallel Q) - \mathrm{JSD}_{B,\alpha/2},$$

where $\mathrm{JSD}_{B,\alpha/2}$ and $\mathrm{JSD}_{B,1-\alpha/2}$ are the $\alpha/2$-th and $1-\alpha/2$-th percentiles of the bootstrap JSD estimates, respectively.

3.1. Temporal Item Distribution

To assess how the two corpus versions differ temporally, we compare the distributions of yearly item frequencies in the original PG split and our fine-grained PG split, with frequencies normalized by the total number of items in each corpus. In addition to the raw yearly frequencies, we compute LOWESS-smoothed frequency curves (with smoothing parameter frac $= 0.2$) to highlight broader temporal trends (see Figure 2). Both corpora are annotated using the publication-year annotation method of Momen et al. (2025) (see Section 2.3 for details of the experimental setup).

3.1.1. Results and Analysis

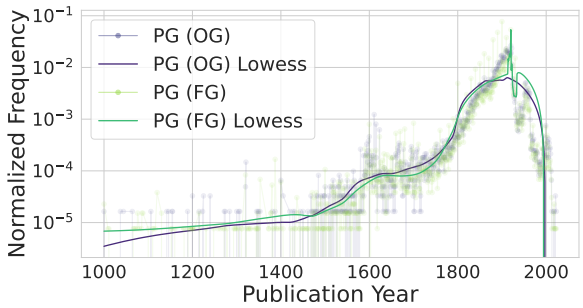

Figure 2: Comparison of yearly item frequencies between the original PG split and our fine-grained PG split, with frequencies normalized by the total number of items in each corpus (OG = original; FG = fine-grained).

From the LOWESS-smoothed distribution of our fine-grained corpus, we observe a more pronounced rise in the number of items leading up to 1920, followed by a sharp decline thereafter, with item counts dropping to roughly 10% of their 1920 level compared to the original PG split. Importantly, divergence decreases monotonically over time, suggesting that structural inconsistencies are concentrated in earlier, sparsely populated corpus sections. We quantify this trend by computing $\mathrm{JSD}_{\text{corrected}}$ over 200-year intervals between 1000 and 2000 (see Table 1). Divergence is highest for early slices (e.g., 1000–1199: 0.587) and lowest for later periods (e.g., 1800–1999: 0.047), with an overall divergence of 0.047. While the aggregate divergence across the full time span is moderate, early periods exhibit substantial divergence (0.59), indicating that coarse collection-level grouping disproportionately distorts sparsely represented historical slices.

Time Slice	JSD	95% CI
1000–1199	0.5867	[0.5067, 0.6800]
1200–1399	0.3798	[0.3058, 0.4508]
1400–1599	0.0654	[0.0422, 0.0910]
1600–1799	0.0574	[0.0495, 0.0667]
1800–1999	0.0467	[0.0455, 0.0480]
full	0.0471	[0.0458, 0.0486]

Table 1: $\mathrm{JSD}_{\text{corrected}}$ and 95% $\mathrm{CI}_{\text{corrected}}$ for 200-Year time slices between the original PG split and our fine-grained split.

3.2. Per-Language Distributional Effects

For the second experiment, we examine whether the fine-grained split captures diachronic linguistic phenomena, such as negation, more distinctly across languages. We include all languages for which D-Neg supports negation detection and that are represented by at least 100 items in the corpus, yielding the following set: English, Italian, French, German, Spanish, Dutch, and Chinese. To this end, for each language we construct a joint empirical distribution over publication year and negation frequency, where negation frequency is computed by first dividing the total number of detected negation cues in an item by the number of sentences in that item and then averaging this value across all items published in the same year. We use the publication-year annotations described in Section 2.3 and the negation annotations described in Section 2.4. We then compute pairwise $\mathrm{JSD}_{\text{corrected}}$ between these language-specific distributions in both corpus splits (see Figure 3). This allows us to assess whether the fine-grained version yields more differentiated and linguistically interpretable cross-linguistic relationships, including similarities and dissimilarities potentially associated with properties such as negative concord or shared language family.

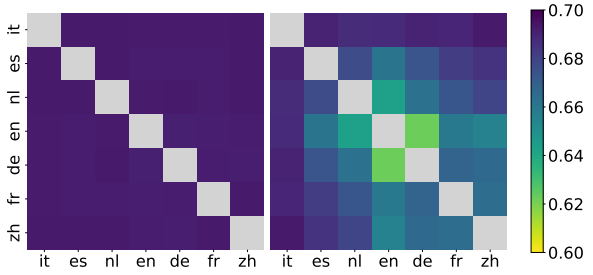

Figure 3: Pairwise $\mathrm{JSD}_{\text{corrected}}$ between language-specific joint distributions over publication year and year-level average negation frequency for the original (left) and fine-grained (right) PG splits.

3.2.1. Results and Analysis

In the original PG split, the language-specific joint distributions of publication year and negation frequency exhibit uniformly high JSD values, suggesting limited differentiation across languages, which may obscure typologically meaningful relationships. In contrast, the fine-grained split reveals more structured variation, including lower divergence among typologically related languages (e.g., Germanic non-concord languages such as English, German, and Dutch). We do not interpret this greater dispersion as inherently better. Rather, we take it to be desirable insofar as it reduces near-uniform divergence patterns that may result from noisy publication-year metadata and yields relationships that are more consistent with independently established typological structure in negation (Haspelmath, 2013). English appears relatively central, with low divergence to most languages, whereas Italian remains more isolated. Italian's more isolated position may partly reflect language-specific negation properties, as Standard Italian has been analyzed as a non-strict negative concord language (Poletto and Oliviéri, 2018), although corpus-compositional factors may also contribute. English, by contrast, may appear more central because its larger corpus representation likely yields a more stable aggregate profile. To quantify dispersion across language pairs, we compute the Interquartile Range (IQR) of the divergence values. The IQR increases from 0.001 in the original split to 0.025 in the fine-grained split, a 25-fold increase, indicating substantially greater cross-linguistic variability. A two-sample Kolmogorov–Smirnov test confirms that the distributions differ significantly ($D = 0.9524$, $p < 0.001$).

4. Related Work

DTA on the PG Corpus. Several studies have applied DTA to the PG corpus. These include investigations of the stylistic evolution of literary authors over time (Klaussner and Vogel, 2018) and their influence (Hughes et al., 2012), the classification of diachronic semantic shifts (Koldenhof, 2021), and detailed analyses of changing emotional or thematic patterns, such as the study of melancholy (Kung, 2007).

PG Publication Year Refinement. To date, only a few studies have attempted to refine the temporal labels of PG texts. Gerlach and Font-Clos (2020) relied on author lifespan estimates, Hegde et al. (2025) augmented these with external sources such as Wikipedia and OpenLibrary, and Momen et al. (2025) combined previous approaches with LLM-based RAG applied to the first and last pages of each item.

5. Conclusion

We address a critical limitation of the widely used Project Gutenberg (PG) corpus that has been largely overlooked: approximately 15% of its entries are collections rather than individual works. This structural issue undermines the validity of diachronic analyses on PG. By automatically detecting and splitting these collections into their atomic items (e.g., books, letters, poems), we produce a more temporally faithful and linguistically coherent resource. Using negation as a case study, we demonstrate that our fine-grained corpus enables more precise and interpretable analyses of diachronic linguistic phenomena. To support further research, we publicly release the fully formatted corpus in UIMA format, including all metadata, publication year annotations, segmentation, and negation annotations. Our work establishes a stronger foundation for large-scale historical language studies and underscores the importance of structural and metadata integrity in computational linguistics.

6. Limitations & Future Work

For publication-year annotation, we adopt the method of Momen et al. (2025) and use the same models as in their original setup. These models are compact and efficient, which is beneficial for a corpus at the scale of Project Gutenberg, although future work could examine whether larger models yield more accurate temporal annotations. Our approach to splitting collections currently relies on an effective heuristic; future work could explore whether this step can be handled more systematically with LLM-based methods. Finally, our analysis focuses on negation. Extending the evaluation to additional linguistic phenomena would help determine how broadly the observed effects generalize.

7. Ethics Statement

Our work builds on the Project Gutenberg corpus, a collection of public-domain texts, and does not involve any sensitive, private, or personal data. All texts used are freely available and legally redistributable, and our annotations and processed corpus are released in compliance with PG's terms of use.

We acknowledge that the PG corpus is biased toward Western, English-language, and literary texts, thus underrepresenting other languages, genres, or historical perspectives. Our fine-grained corpus does not mitigate these intrinsic corpus biases; users of Gutenberg+ should remain aware of these limitations when drawing linguistic or cultural conclusions.

8. Acknowledgements

This work was supported by the German Research Foundation (DFG) as part of Project INF within CRC 1629 "Negation in Language and Beyond" (NegLaB - https://www.neglab.de/).

9. Bibliographical References

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André F T Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. EuroBERT: Scaling Multilingual Encoders for European Languages. In *Second Conference on Language Modeling*, pages 1–28, Montreal, Canada.

William Croft. 1991. The evolution of negation. *Journal of Linguistics*, 27(1):1–27.

Bradley Efron and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

David Ferrucci and Adam Lally. 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1).

Leon Lukas Hammerla, Andy Lücking, Carolin Reinert, and Alexander Mehler. 2025. D-neg: Syntax-aware graph reasoning for negation detection. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1432–1454, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Martin Haspelmath. 2013. Negative indefinite pronouns and predicate negation (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Niharika Hegde, Subarnaduti Paul, Lars Joel-Frey, Manuel Brack, Kristian Kersting, Martin Mundt, and Patrick Schramowski. 2025. Chronoberg: Capturing language evolution and temporal awareness in foundation models.

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

James M. Hughes, Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686.

Otto Jespersen, Brett Reynolds, Peter Evans, and Olli O. Silvennoinen. 2025. *Negation in English and other languages*. Number 8 in Classics in Linguistics. Language Science Press, Berlin.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Carmen Klaussner and Carl Vogel. 2018. A diachronic corpus for literary style analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dylan Koldenhof. 2021. Word embeddings to classify types of diachronic semantic shift. University of Twente.

Sally Kung. 2007. A diachronic study of melancholy in a british novel corpus. University of Birmingham.

Gabriella Mazzon. 2016. *A History of English Negation*. Routledge.

Omar Momen, Manuel Schaaf, and Alexander Mehler. 2025. Filling the temporal void: Recovering missing publication years in the Project Gutenberg corpus using LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17318–17334, Vienna, Austria. Association for Computational Linguistics.

Cecilia Poletto and Michèle Oliviéri. 2018. *Chapter 9. Negation patterns across dialects*, pages 133–148. John Benjamins Publishing Company.

10. Language Resource References

Project Gutenberg. 1971-2026. *Project Gutenberg Corpus*. Project Gutenberg. Open-access digital library; available at https://www.gutenberg.org, accessed August 22, 2025.