Modeling the Invisible: Applying the Unseen Species Model to Chivalric Literature in the Iberian Peninsula

Carolina Macedo[1,2]

[1] École nationale des chartes, Paris, France
[2] Biblissima, cluster 7, Paris, France

Abstract

This study applies an unseen species model—a non-parametric estimator originally developed in ecology—to the problem of textual loss in medieval Iberian chivalric literature. Drawing on abondance data for Castilian, Portuguese, and Catalan works, it estimates the number of lost or unrecorded texts based on the frequency of rare items. The results suggest that only about 40 % of the original corpus and roughly 8 % of individual documents have survived, indicating that the extant record represents only a small fraction of the field's former scale. Beyond numerical estimation, the approach demonstrates how probabilistic modelling can inform literary historiography, providing a structured framework for reasoning about absence and survival.

Keywords: chivalric literature, Iberian Peninsula, manuscript studies, unseen species model, textual loss, digital humanities, computational philology

1 Introduction

As with much of the historical record, medieval literature survives only in fragments, inevitably shaping a partial and biased understanding of the whole. Chivalric literature—a cornerstone of the medieval European imagination—played a vital role in shaping cultural memory and transmitting heroic ideals across generations. Widely copied, translated, and adapted, these narratives circulated broadly throughout the late Middle Ages, yet their manuscript transmission was uneven. Much of this corpus has been lost or survives only in fragmentary form, especially in the Iberian Peninsula, where French originals were early translated into Castilian, Catalan, and Portuguese. Scholars have long recognized this uneven preservation, but the magnitude of the loss—and its implications for literary history—remain difficult to quantify. The richness of Iberian chivalric production contrasts sharply with the fragility of its material record: most works are known from only a few manuscripts, many from none at all.

Recent years have seen growing use of quantitative approaches in the study of cultural and philological domains, offering new ways to address problems of incompleteness and loss. Among these, *unseen species models*—statistical tools originally developed in ecology to estimate biodiversity from partial samples—have proved especially promising. Designed to infer the number of unobserved species from the distribution of rare ones [8; 9; 27], such models have since been adapted to different fields including literary studies. Kestemont *et al.* [30; 32] applied this framework in their study *Forgotten Books*, estimating the proportion of lost medieval works from the frequency of rarely attested texts.

Building on this approach, the present article applies an unseen species model—specifically the Chao1 estimator—to a corpus of Iberian chivalric works in Castilian, Portuguese, and Catalan.

Carolina Macedo. "Modeling the Invisible: Applying the Unseen Species Model to Chivalric Literature in the Iberian Peninsula." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1354–1363. https://doi.org/10.63744/qH01jSZULykB.

It asks two related questions: (1) How extensive was textual loss within these traditions? and (2) What do these estimates reveal about the dynamics of preservation and disappearance across linguistic and institutional contexts?

Treating the surviving corpus as a statistical sample rather than a fixed canon aligns with recent perspectives in evolutionary cultural studies that view transmission as a process shaped by selection, bias, and contingency [1; 25]. The Iberian chivalric corpus thus serves as a test case for how probabilistic reasoning can complement historical interpretation—revealing not only what has been preserved, but also what patterns of survival disclose about the mechanisms of cultural endurance and forgetting.

2 Corpus

2.1 Corpus Construction

The corpus[1] considered here brings together works of chivalric and heroic literature transmitted in manuscript form and produced or translated across the main vernacular traditions of the Iberian Peninsula—primarily Castilian, Catalan, and Portuguese (or Galician-Portuguese)—between the 13- 16th centuries. The selection builds upon manuscript material identified in existing scholarship on Iberian romance literature (e.g., [2; 3; 4; 5; 7; 14; 16; 34; 38]), as well as specialised bibliographic databases and dictionaries [24; 35; 37] and dedicated textual corpora [13]. Thematically, the study adopts the notion of *materia caballeresca*, which encompasses not only traditional chivalric romances but also narrative texts that integrate heroic, epic, or hagiographic elements within a chivalric framework. This concept denotes a shared narrative imaginary rooted in medieval ideals of knighthood, adventure, and moral exemplarity, shaped by both military and courtly values and informed by codes of conduct that unite martial, social, and spiritual dimensions [26]. The corpus therefore comprises both original Iberian compositions—such as the *Libro del caballero Zifar* and the *Amadís de Gaula* (in its earliest version)—and medieval translations or adaptations of Arthurian material and other established chivalric cycles, including the *Baladro del sabio Merlín* and the *Demanda do Santo Graal*. It also encompasses the five hagiographic-chivalric texts contained in manuscript h-i-13 and the verse epic tradition, within which certain works of the *mester de clerecía* exhibiting a chivalric dimension, such as *El Libro de Alexandre*. These examples are illustrative rather than exhaustive, highlighting the generic and thematic range of the corpus.

2.2 Corpus Delimitation

We exclude from this corpus the so-called *libros de caballerías*, even those that survive in manuscript form, as they are protagonists of the new, predominantly Castilian, print culture. They belong to a transitional phase of the genre between the late medieval and early modern periods, characterized by print transmission and a distinct literary culture. Most of the surviving manuscript witnesses are copies of printed editions—drafts or fair copies preserved in libraries without wide circulation, or professional and personal copies. These late *libros de caballerías* manuscripts thus reproduce already printed material and participate in a fundamentally different mode of textual transmission. Consequently, this study excludes this group, along with other subgenres that emerged in close connection with print culture—such as the *historias caballerescas breves* and the *ficción sentimental*. These exclusions reflect not only the historical boundaries of the corpus but also the methodological distinctions required by our analytical framework. The dynamics of print transmission differ significantly from those of manuscript culture, and the methods applied to computational modeling likewise vary according to the nature of the material. This delimitation

[1] For a detailed overview of the corpus, see the CSV file, which includes information on titles, authors, languages, dates of production, and the current locations of the manuscripts.

allows the study to focus exclusively on the medieval formation of chivalric narrative and its transmission within the manuscript tradition, prior to the transformations introduced by print.

Also, since this study focuses on observable data—documents that are materially attested, either as complete *codices* or identifiable fragments—texts known only through titles, summaries, or indirect mentions[2] were not included in the corpus,[3] which is therefore limited to works with surviving material witnesses. Although such references illuminate the cultural orientations of the ruling classes, they must be treated with caution: catalogues and inventories rarely reflect the full range of works in circulation.[4] Often compiled as records of cultural treasures, these lists are vague, omitting titles or quantities, and most have not survived the passage of time, as Bogdanow [5] observes. Such material therefore belonged to a protected environment that favored—if only temporarily—their preservation, raising questions about the representativeness it conveys. Classical approaches to estimating the corpus of lost works [18; 19; 20; 21] rely heavily on this evidence, a method that risks bias by privileging institutionally preserved texts and underestimating the number and diversity of works circulating beyond such frameworks. By contrast, the approach adopted here grounds the model in verifiable data while allowing for the statistical estimation of additional works that may have existed but are no longer preserved in the surviving record.

2.3 Corpus Overview

Although the transmission of medieval chivalric texts within the Iberian context reveals distinct tendencies among the various vernacular traditions, several striking commonalities nevertheless emerge. The study of these traditions invites a reconsideration of key aspects of the preservation and loss of such works. In this regard, we have examined several features of the texts in our corpus—most notably their fragmentary condition, their generally late production, the predominance of Castilian manuscripts over those in other Iberian languages, and the comparatively fragile Portuguese tradition, preserved almost exclusively through unique manuscript witnesses.[5]

While the chivalric romance genre is relatively well represented across the three traditions, the surviving evidence for the Iberian epic tradition is extremely scarce and, at least for now, limited to Castilian. This material absence has even led some researchers to doubt whether an epic tradition ever existed in Portuguese, for example. Even so, its narratives continued to circulate within other genres, particularly chronicles and later *romanceros*. Some scholars have gone so far as to suggest that this now-lost corpus could be partially reconstructed—at least thematically—through the intertextual traces preserved in these later works [17], whose authors, over time, reformulated or suppressed parts of the original narratives according to the tastes and conventions of their own age.

The corpus records 23 distinct works—each an immaterial entity upon which the text depends—and 45 surviving manuscript witnesses that materialize and transmit those works within the Iberian chivalric tradition. Most of the works are anonymous, and prose constitutes the predominant textual form. Chronologically, the witnesses range from the thirteenth to the sixteenth century, with a marked concentration in the fifteenth. The corpus is largely dominated by the *romance de cavalaria* genre (21 occurrences), which accounts for nearly half of the surviving manuscripts. Castilian is by far the most represented language (78.57%), followed by Portuguese

[2] The main sources of such indirect references include royal and ecclesiastical library catalogues, allusions in prologues or colophons of other works, and mentions preserved in last wills and household inventories.

[3] Nor were those manuscripts found and subsequently lost again in recent decades, such as the fragments of the *Lancelot* in Catalan: two folios from the mid-fourteenth century formerly held in the private library of Francesc Cruzate of Mataró.

[4] As Sousa Viterbo [39] notes, these lists were far from bibliographical catalogues in the modern sense: they emphasized the monetary and artistic value of books—particularly their bindings—rather than their intellectual content.

[5] These aspects are discussed in greater detail in Chapter 3, “Trajectoire des manuscrits : aperçus préliminaires sur leur conservation et leur disparition,” of my MA thesis [36].

and Catalan—proportions that align with the broader linguistic and cultural patterns observable in the Iberian chivalric corpus.

3 The Unseen Species Model and Its Application to the Corpus

Originally developed in ecology to estimate species richness from incomplete samples [8; 9; 10], unseen species models have since found broad application across diverse domains [22; 23]. More recently, they have been applied to the study of cultural and literary phenomena [29; 30; 32; 33]. In this context, the model enables researchers to estimate the probable number of lost or unobserved works by analyzing the frequency of rare items within a known corpus. To implement the model, each chivalric work was treated as a unique narrative item, and its frequency was defined by the number of distinct material witnesses in which it is attested. Within the unseen species framework, each work corresponds to a single "species," while each material attestation—whether complete or fragmentary—constitutes an "observation" of that species.

The key insight of the model is that a sample containing many rare items—particularly singletons (attested once) and doubletons (attested twice)—is statistically more likely to be incomplete [12; 27]. The distribution of these rare items thus serves as a proxy for estimating the number of undetected species (here unobserved works), thereby approximating the original species richness. Several estimators are available to researchers for assessing species diversity. The most widely used methods for estimating species richness include the Chao estimators—named after biostatistician Anne Chao (Chao1 for *abundance data* and Chao2 for *incidence data*)—and the jackknife estimator [15; 28]. Chao1 is one of the most commonly used estimators for measuring species richness, as this richness can neither be precisely quantified nor directly estimated through observation. This nonparametric lower-bound estimator can be defined as follows (e.g., [9; 10]):

$$\hat{S}_{\text{Chao1}} = S_{\text{obs}} + \frac{f_1^2}{2f_2}$$

where:

- S_{obs} is the number of observed works in the dataset,
- f_1 is the number of singletons,
- f_2 is the number of doubletons.

The Chao1 estimator was applied to frequency data (our *abondance data*) derived from the corpus described above. Although inevitably incomplete, the corpus internal frequency distribution allows for a statistically meaningful analysis. Rather than offering definitive totals, the results provide lower-bound estimates that complement traditional literary historiography and help quantify the scale of cultural loss in pre-modern textual traditions. This model is particularly well-suited to medieval textual traditions, which are often transmitted unevenly and preserved only in small, fragmentary datasets. It treats the observed corpus as a sample drawn from a larger population of original works, many of which may no longer survive.

The statistical computation was performed using the `copia` Python package developed by Kestemont and Karsdorp [31]. This package implements the Chao1 estimator and other species richness estimators commonly used in ecology and cultural analytics.

To assess the robustness of the approach, the model was applied to the complete corpus as well as to distinct linguistic subcorpora, allowing comparison across traditions with varying levels of attestation. For each language group (Castilian, Catalan, Portuguese), frequency distributions were compiled, and the counts of singletons (f_1) and doubletons (f_2) were extracted as model inputs.

4 Results

The application of the Chao1 estimator to the Iberian chivalric corpus reveals a significant degree of textual loss across all three linguistic traditions. The results presented below combine raw frequency data with statistical estimations and interpretative insights (developed in section 5), highlighting both the potential and the limits of the method.

4.1 Corpus Statistics

Table **??** summarizes the observed values for each subcorpus; these constitute our *abondance data*. Portuguese and Catalan traditions present extremely small samples, with only 4 and 5 works respectively, and very few doubletons. Castilian, while somewhat larger, still reflects a high proportion of rare items (16 singletons out of 21 works), suggesting fragility in transmission.

4.2 Estimated Richness

In table 2 the Chao1 estimator suggests that the actual number of distinct works is significantly higher than what survives. In Portuguese, the estimate nearly doubles the observed count. In Catalan, the estimate also doubles, indicating that for every known text, another likely existed. The Castilian tradition shows the largest gap, with a potential total of over 60 works—three times the number observed. These confidence intervals are extremely wide, especially in the Castilian case, where the upper bound exceeds 170. This may reflect the model's sensitivity to skewed data and the sparsity of multi-attested works. Despite this uncertainty, reflected in the wide confidence intervals, the estimates suggest a major loss of textual material. For Portuguese, the lower confidence bound falls slightly below the number of observed works ($S = 4$) due to sampling variance. In practice, however, the true number of works cannot be smaller than the observed richness, so this lower bound should be interpreted as a statistical artefact.

4.3 Combined Corpus and Survival Rates

When all three traditions are pooled into a single Iberian corpus (table 3), the estimations remain more consistent: Chao1 suggests around 54 works once existed, versus 23 observed. For documents, the gap is even starker: from 44 known attestations, the model extrapolates nearly 500 potential documents. From these estimates, we derive survival rates: approximately 40% of works and only 8% of documents are currently known. The implied loss—60% of works and over 90% of manuscripts—is striking. This confirms the fragile nature of medieval literary transmission, especially for less-institutionalised or linguistically minor traditions.

5 Discussion

The application of the unseen species model to the Iberian chivalric corpus offers a new perspective on a long-standing problem in literary history: the extent and impact of textual loss. While the fragmentary nature of medieval transmission has long been acknowledged, its magnitude has remained elusive. The estimates presented here do not recover missing works, but they provide a structured way to think about what has been lost—and what patterns this loss may reveal.

One striking result is the high proportion of singletons in all three small subcorpora, which strongly influences the estimates of unseen works [10]. This prevalence suggests a corpus characterized by instability: many works may have circulated in a very limited number of copies, increasing their vulnerability to disappearance. It also invites reflection on how textual survival is conditioned not only by production but by preservation, copying, storage, and cataloguing—processes that are historically contingent and unevenly distributed across regions and institutions.

The overall abundance of surviving documents, even when all languages are considered, remains very modest. This scarcity, combined with the predominance of singletons (f_1), substantially increases the likelihood of loss. As Kestemont *et al.* [32] observed for the French corpus, such long-tailed distributions make even large literatures vulnerable to *immaterial loss*. For comparative purposes, we referred to the estimation plots of the languages analysed in *Forgotten Books*, which highlight how the method behaves when applied to larger samples. Notably, the distribution for Castilian, while not identical, more closely resembles that of smaller literary traditions such as Icelandic or Irish within that corpus.

The Iberian tradition thus presents several distinctive features when compared with other European contexts. In contrast to the vast and relatively well-preserved French corpus, Iberian chivalric literature survives in a far more fragmentary state. Within the Peninsula, the Portuguese and Catalan areas appear to have suffered particularly severe losses; paradoxically, however, the Portuguese corpus preserves the highest proportion of complete codices among all the traditions. Another distinctive feature is the late persistence of manuscript production: the habit of copying chivalric texts by hand extended well beyond the advent of print, reflecting the longevity of medieval scribal practices in the Iberian literary sphere — and carrying with it all the fragility inherent to manuscript transmission.

In *Forgotten Books*, Kestemont *et al.* argue that evenness—the balanced distribution of elements within a system—enhances stability in both ecological and cultural contexts [11]. Just as uniform ecosystems tend to be more resilient to disruption, literary traditions with a more even distribution of works, such as Irish and Icelandic literatures, appear better preserved than larger, more uneven continental canons. The Iberian case, however, seems to illustrate this latter pattern more clearly. Despite its relative geographical isolation—almost insular in form—the region displays relative unevenness, and thus seems to have been less shielded from disruptive forces. Political fragmentation, linguistic diversity, and uneven institutional development may have weakened the mechanisms of textual preservation that, in more uniform traditions, foster resilience. In the Iberian context, unevenness appears to have amplified vulnerability, exposing smaller or less institutionalised literary traditions, such as the Catalan corpus, to a higher risk of loss.

From a methodological standpoint, the use of the Chao1 estimator demonstrates the potential of computational tools for literary historiography. It highlights the value of thinking probabilistically about the corpus—treating what survives as a sample rather than as a canon. Such an approach invites us to reconsider the notion of a stable canon and, instead, to think in terms of an ephemeral canon—one shaped as much by absence as by presence. What survives may reflect not intrinsic aesthetic or cultural value, but rather the contingencies of archival preservation, copying practices, institutional patronage, or simple randomness [6].

6 Conclusion

This study applied an unseen species model to a corpus of Iberian chivalric works in Castilian, Portuguese, and Catalan, estimating the extent of textual loss using frequency-based statistical methods. The results suggest that only about 40% of original works and 8% of documents have survived, pointing to a much larger literary field than the current corpus reflects. While the estimates cannot recover lost works, they provide a principled means of reasoning about what remains unknown and about the conditions that shape what survives.

This study demonstrates how computational approaches can contribute to literary historiography. Rather than replacing traditional philology, this approach complements it by offering a quantitative lens on cultural memory and loss. Quantitative modelling thus offers new perspectives on the gaps and silences of literary history. By adapting a model from ecology to the study of medieval literature, this work contributes to a broader rethinking of what it means to work with incomplete corpora—where absence, approached with the right tools, becomes a form of evidence.

This paper represents only a preliminary outline of a larger project that remains to be developed. Future research may refine these estimates through a more precise delineation and analysis of the corpus, drawing on existing metadata on manuscript transmission and preservation to reveal additional patterns of survival.

Acknowledgments

This article expands upon research initially conducted as part of a master's thesis.[6] The study also builds on a computational approach recently proposed by Kestemont *et al.* [30; 32], specifically in their article *Forgotten Books*. The author wishes to thank the anonymous reviewers for their insightful comments and suggestions, which greatly contributed to improving this paper.

References

[1] Acerbi, Alberto, Mesoudi, Alex, and Smolla, Marco. *Individual-Based Models of Cultural Evolution: A Step-by-Step Guide Using R*. 1st ed. London: Routledge, May 2022. ISBN: 978-1-00-328206-8. DOI: 10.4324/9781003282068.

[2] Ailenii, Simona. *Os primeiros testemunhos da tradução galego-portuguesa do romance arturiano*. PhD thesis. Porto: Universidade do Porto, Faculdade de Letras, 2012.

[3] Alvar, Carlos and Gómez Moreno, Ángel. *La Poesía Épica y de Clerecía Medievales*. Vol. 2. Historia Crítica de La Literatura Hispánica. Madrid: Taurus, 1988.

[4] Balaguer, Pere Bohigas i. "Un Nou fragment del "Lançalot" català". In: *Estudis Romànics 10* (1967), pp. 179–187. ISSN: 2013-9500.

[5] Bogdanow, Fanni. *La Version Post-Vulgate de La Queste Del Saint Graal et de La Mort Artu*. Vol. 4 vols. Société des Anciens Textes Français, 1991.

[6] Camps, Jean-Baptiste and Randon-Furling, Julien. "Lost Manuscripts and Extinct Texts: A Dynamic Model of Cultural Transmission". In: *Proceedings of the Computational Humanities Research Conference 2022 Antwerp,Belgium, December 12-14, 2022, Anvers*. CEUR Workshop Proceedings 3290. Anvers, 2022, pp. 198–214. DOI: 10.48550/arxiv.2210.16577.

[7] Castro, Ivo. "Sobre a Data Da Introdução Na Península Ibérica Do Ciclo Arturiano Da Post-Vulgata". In: *Boletim de Filologia* , no. XVIII (1983), pp. 81–98.

[8] Chao, Anne. "Nonparametric Estimation of the Number of Classes in a Population". In: *Scandinavian Journal of Statistics 11*, no. 4 (1984), pp. 265–270. ISSN: 0303-6898. JSTOR: 4615964.

[9] Chao, Anne and Chiu, Chun-Huo. "Species Richness: Estimation and Comparison". In: *Wiley StatsRef: Statistics Reference Online*, ed. by Ron S. Kenett, Nicholas T. Longford, Walter W. Piegorsch, and Fabrizio Ruggeri. 1st ed. Wiley, Aug. 2016, pp. 1–26. ISBN: 978-1-118-44511-2. DOI: 10.1002/9781118445112.stat03432.pub2.

[10] Chao, Anne, Chiu, Chun-Huo, Colwell, Robert K., Magnago, Luiz Fernando S., Chazdon, Robin L., and Gotelli, Nicholas J. "Deciphering the Enigma of Undetected Species, Phylogenetic, and Functional Diversity Based on Good-Turing Theory". In: *Ecology 98*, no. 11 (Nov. 2017), pp. 2914–2929. ISSN: 0012-9658, 1939-9170. DOI: 10.1002/ecy.2000.

[11] Chao, Anne and Ricotta, Carlo. "Quantifying Evenness and Linking It to Diversity, Beta Diversity, and Similarity". In: *Ecology 100*, no. 12 (Dec. 2019), e02852. ISSN: 0012-9658, 1939-9170. DOI: 10.1002/ecy.2852.

[6] See the corresponding GitHub repository that hosts the complete research project.

[12] Chao, Anne et al. “Quantifying Sample Completeness and Comparing Diversities among Assemblages”. In: *Ecological Research 35*, no. 2 (2020), pp. 292–314. ISSN: 1440-1703. DOI: 10.1111/1440-1703.12102.

[13] Corfis, Ivy and Ancos, Pablo. “CHCR: Corpus of Hispanic Chivalric Romances”. 2005. URL: https://textred.spanport.wisc.edu/chivalric/texts.html.

[14] Cuesta Torre, María Luzdivina. “La transmisión textual de ”Don Tristán de Leonís””. In: *Revista de literatura medieval 5* (1993), pp. 63–93. ISSN: 1130-3611.

[15] Daly, Aisling J., Baetens, Jan M., and De Baets, Bernard. “Ecological Diversity: Measuring the Unmeasurable”. In: *Mathematics 6*, no. 7 (July 2018), p. 119. ISSN: 2227-7390. DOI: 10.3390/math6070119.

[16] de Almeida Toledo Neto, Sílvio. “Os testemunhos portugueses do Livro de José de Arimatéia e o seu lugar na tradição da Estoire del Saint Graal: colação de exemplos”. In: *De Cavaleiros e Cavalarias. Por terras de Europa e América*. São Paulo: Humanitas, 2012, pp. 579–589.

[17] Deyermond, Alan D. *Epic Poetry and the Clergy: Studies on the ” Mocedades de Rodrigo”*. London: Tamesis Books, 1968.

[18] Deyermond, Alan D. *La literatura perdida de la Edad Media castellana: catálogo y estudio*. Obras de referencia 7. Salamanca: Ed. Universidad de Salamanca, 1995.

[19] Deyermond, Alan D. “Lost Literature in Medieval Portuguese”. In: *Medieval and Renaissance Studies in Honour of Robert Brian Tate*. 1986, pp. 1–12.

[20] Deyermond, Alan D. “The Lost Genre of Medieval Spanish Literature”. In: *Hispanic Review 43* (1975), pp. 231–259.

[21] Deyermond, Alan D. “The Problem of Lost Epics: Evidence and Criteria”. In: *Olifant 6*, no. 1 (1978), pp. 35–38. JSTOR: 45297872.

[22] Efron, Bradley and Thisted, Ronald. “Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?” In: *Biometrika 63*, no. 3 (1976), pp. 435–447. DOI: 10.1093/biomet/63.3.435.

[23] Eren, Metin I., Chao, Anne, Hwang, Wen-Han, and Colwell, Robert K. “Estimating the richness of a population when the maximum number of classes is fixed: A nonparametric solution to an archaeological problem”. In: *PLOS ONE 7*, no. 5 (2012), e34179. DOI: 10.1371/journal.pone.0034179.

[24] Faulhaber, Charles B. “PhiloBiblon”. University of California. Berkeley, 1997. URL: https://philobiblon.upf.edu/html/index.html.

[25] Gabora, Liane and Aerts, Diederik. “Distilling the essence of an evolutionary process, and implications for a formal description of culture”. In: *Proceedings of Center for Human Evolution Workshop #5: Cultural Evolution, May 2000, Foundation for the Future, Seattle WA*, ed. by W. Kistler. W. Kistler / Foundation for the Future, 2005. URL: https://arxiv.org/abs/1309.4712.

[26] Gómez Redondo, Fernando. “La literatura caballeresca castellana medieval”. In: *Amadís de Gaula, 1508: quinientos años de libros de caballerías*. Madrid: Biblioteca Nacional de España and Sociedad Estatal de Conmemoraciones Culturales, 2008, pp. 53–79.

[27] Good, I. J. “The Population Frequencies of Species and the Estimation of Population Parameters”. In: *Biometrika 40*, no. 3-4 (1953), pp. 237–264. DOI: 10.1093/biomet/40.3-4.237.

[28] Gotelli, N. and Colwell, Robert. “Estimating Species Richness”. In: *Frontiers in Measuring Biodiversity*. Vol. 12. Oxford University Press, Jan. 2011, pp. 39–54.

[29] Karsdorp, Folgert, Wevers, Melvin, and Lottum, Jelle van. “What Shall We Do with the Unseen Sailor? Estimating the Size of the Dutch East India Company Using an Unseen Species Model.” In: *Proceedings of the Computational Humanities Research Conference, 2022, 189–97. Antwerp, Belgium*. CEUR Workshop Proceedings 3290.

[30] Kestemont, M. and Karsdorp, Folgert. “Estimating the Loss of Medieval Literature with an Unseen Species Model from Ecodiversity”. In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020) Amsterdam, the Netherlands, November 18-20, 2020.* CEUR Workshop Proceedings 2723. 2020, pp. 44–55.

[31] Kestemont, Mike and Karsdorp, Folgert. “Copia: Bias correction for richness in abundance data”. 2022. URL: `https://github.com/mikekestemont/copia`.

[32] Kestemont, Mike, Karsdorp, Folgert, De Bruijn, Elisabeth, Driscoll, Matthew, Kapitan, Katarzyna A., Ó Macháin, Pádraig, Sawyer, Daniel, Sleiderink, Remco, and Chao, Anne. “Forgotten Books: The Application of Unseen Species Models to the Survival of Culture”. In: *Science (New York, N.Y.) 375*, no. 6582 (Feb. 2022), pp. 765–769. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.abl7655`.

[33] Koeser, Rebecca S. and LeBlanc, Zachary. “Missing Data, Speculative Reading”. In: *Modernism/modernity 9*, no. 2 (2024). DOI: `10.26597/mod.0298`. URL: `https://doi.org/10.26597/mod.0298`.

[34] Lucía Megías, José Manuel. “Literatura Caballeresca Catalana: De Los Testimonios a La Interpretación (Un Ensayo de Crítica Ecdótica)”. In: *Caplettra* , no. 39 (2005), pp. 231–256.

[35] Lucía Megías, José Manuel and Alvar Ezquerra, Carlos. *Diccionario Filológico de Literatura Medieval Española. Textos y Transmisión*. Madrid: Castalia, 2002.

[36] Macedo, Carolina. *Le modèle des espèces non vues appliqué à la littérature chevaleresque dans la Péninsule Ibérique*. Available online at *Academia.edu*. MA thesis. Paris: École nationale des chartes, Université PSL, 2024.

[37] Ricardo Pichel Gotérrez and Esther Corral Díaz, edited by. *Guía Para o Estudo Da Prosa Galega Medieval*. Santiago de Compostela: Xunta da Galicia, 2021.

[38] Sharrer, Harvey L. *A Critical Bibliography of Hispanic Arthurian Material. I Texts: The Prose Romance Cycles*. Research Bibliographies & Checklists 3. London: Grant &Cutler Ltd, 1977.

[39] Sousa Viterbo, Joaquim. *A Livraria Real Especialmente No Reinado de D. Manuel : Memoria apresentada á Academia Das Sciencias de Lisboa*. Lisboa: Typ. da Academia, 1901.

A Supplementary Tables

Language	f_1	f_2	S	n	Repositories
Portuguese	3	1	4	5	5
Castilian	16	3	21	33	11
Catalan	4	1	5	6	6

Table 1: Observed frequency statistics by language. f_1 = singletons, f_2 = doubletons, S = number of observed works, n = total number of observations, and "Repositories" indicates the number of distinct source collections.

Language	**Estimate**	**Lower CI**	**Upper CI**
Portuguese	7.6	3.4	14.4
Castilian	62.3	20.4	176.2
Catalan	11.6	5.2	20.7

Table 2: Estimated number of distinct works using Chao1

Category	**Estimate**	**Lower CI**	**Upper CI**
Works	54.2	24.4	146.4
Documents	496.8	56.5	2074.7

Table 3: Estimated richness for the aggregated Iberian corpus