Diversity, survival and bias

Estimating the loss of medieval texts and text carriers using methods from ecodiversity

Mike Kestemont & Folgert Karsdorp

Paper abstract for DHBenelux 2020, 3rd - 5th June 2020

Introduction and state of the art

One of the major hurdles in the scientific study of human cultures in the past is the historic loss of sources. This is no different for historical literary studies: over the centuries, a considerable share of handwritten documents, such as codices or rolls, have been gradually lost as a result of deliberate destruction (e.g., libraries disposing of doubletons) or infrastructural disasters (e.g., library fires). As such, the available data at present only constitute a very limited sample of an original population of literature that was much larger and diverse than the extant materials might suggest. This partial observability is at the heart of this contribution, in which we focus on medieval literature.

Estimates as to the scope and nature of the original population of medieval literature serve an important role in medieval studies [Van Oostrom 2006], which are, for obvious reasons, strongly biased towards the texts and documents that are currently still known to us. For decades, the issue of “lost books” has haunted the field. In codicology and book history [Buringh 2011], surviving descriptions of medieval library holdings can be used, to some extent, to estimate the losses which these collections sustained over the centuries. In the same spirit, book historians [Egghe & Proot 2007] have applied innovative statistical models to early modern prints to gauge how much of the material specimens might have been lost.

The aforementioned studies are concerned with the loss of books (or rather: text carriers), and not with the loss of texts or "works". In this context, textual witnesses (e.g., manuscripts) should be firmly distinguished from the texts that they contain, since medieval texts were regularly (re)copied into multiple, parallel text carriers. At the level of texts too, scholars have advanced hypotheses regarding the number of medieval works that must have disappeared and, thus, the relative loss of textual diversity which it engendered [Van Oostrom 2006]. The existing estimates regarding the loss of texts are often even more informal and “shaky” than those for manuscripts, due to the lack of a suitable, quantitative or empirical framework.

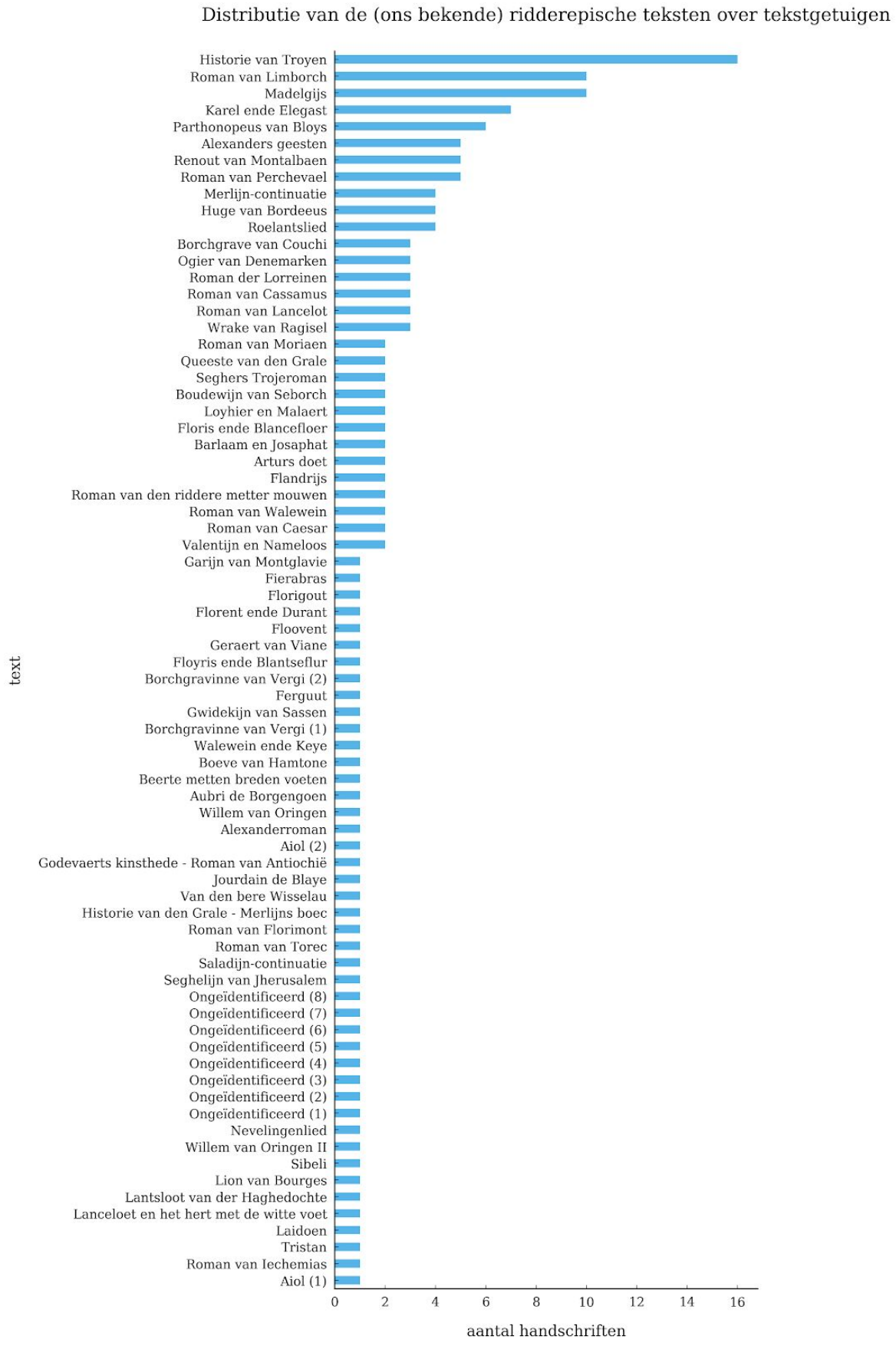

Figure 1: Distribution of surviving Middle Dutch chivalric texts over text carriers.

Ecodiversity

In this paper, we argue that the loss of medieval texts and text carriers can be better estimated using quantitative models from ecodiversity. Estimating the ecological diversity (e.g., the number of distinct species) of a demarcated geographical area is a crucial task in environmental studies and allows us to assess, for instance, the impact of natural disasters on wildlife and other biota [Dale et al. 2018]. During field campaigns, which are by necessity limited in time and staffing, however, it is unlikely that all different species living in that area are actually observed. Statistical methods have been developed to correct the biases in observed counts and estimate the *true* number of species.

Here, we adopt the non-parametric method *Chao1* [Chao & Jost 2015] which can estimate the asymptotic species richness (and a confidence interval, via a bootstrapped procedure). We apply this method to the case study of Middle Dutch chivalric epics and the (fragmentary) sources in which they survive [Kienhorst 1988]. We explicitly build on the analogy that a textual witness can be viewed as a “sighting” of a text, much like an observation of a wildlife species during a sampling campaign in ecology. This method estimates (Fig. 2) that the surviving 74 texts in this epic variety are a sample that was drawn from an original population of ~148 texts. The CI-interval resulting from the bootstrapped procedure, [~106, ~222], reveals that these estimations are generally higher than the more conservative suggestions that have been advanced in conventional philology [Van Oostrom 2006].

Next, we apply an extension of the *Chao1* method [Chao et al. 2009]: after calculating the hypothetical richness of a species assemblage (such as the genre considered here), we, in principle, know the number of species that have not been sighted yet (148 - 74 = 74 texts). This naturally calls into question how much additional manuscripts would then have to be “captured”, in order to observe all of the unsighted species at least once. In Fig. 2, one can think of this value as the number of observations we would need to reach on the X axis, to reach the point where the asymptote starts to saturate (on the Y axis). Here, we assume that this number is a useful proxy for the number of manuscripts that have been lost. This procedure estimates that the original corpus consisted of ~1,952 manuscripts (of which only 164 survived). While this analogy has its limitations, the resulting estimates are in the range of previous estimates from codicologists regarding book loss, i.e. a survival rate of 7% for the category of non-illustrated manuscripts that is best represented in the kind of chivalric epic considered here [Van Oostrom 2006].

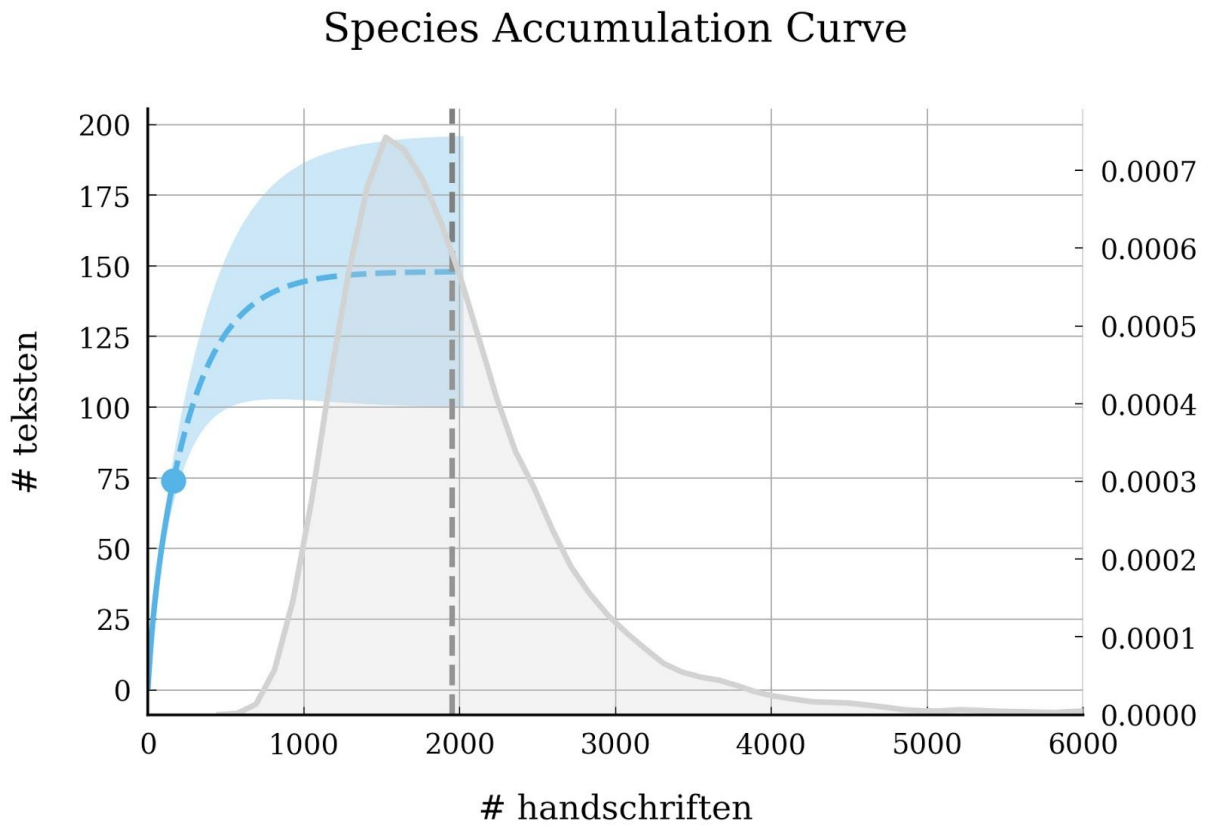

Figure 2: Bootstrapped estimates of the original number of chivalric texts (asymptotic, blue dotted line) and the original number of manuscripts containing them (grey line).

Conclusion

This small-scale case study on Middle Dutch chivalric epics, serves both as a proof of concept and as impetus for further research into the applicability of diversity estimation methods from ecology to the field of medieval literature. The applied methods suggest a survival rate of ~50.00% for the texts in this corpus and ~8.40% for the manuscripts carrying them. Although, in both cases, one should take into account a relatively wide confidence interval, these numbers offer an interesting validation of previous estimates in the field. These methods are especially worthwhile because they rely on a wholly different kind of data than the one traditionally used to gauge medieval book loss. Nevertheless, a number of important issues remain. One primary question is whether the survival of medieval manuscripts can indeed be modeled as a process of *random and independent* sampling. *Chao1* might be non-parametric but fundamentally assumes such a stochasticity in sampling, which we know to correspond only partially to the survival process of medieval books. Texts in miscellanies and convolutes, for instance, are characterized by higher survival rates. Finally, future research should also include a more diverse selection of languages, literatures and repertories. We openly provide the data and code which is necessary to replicate our findings.

References

E. Buringh, *Medieval Manuscript Production in the Latin West. Explorations with a Global Database*. Brill, 2011.

A. Chao & L. Jost, ‘Estimating diversity and entropy profiles via discovery rates of new species’, in: *Methods in Ecology and Evolution* 6 (2015), 873-882.
A. Chao et al., 'Sufficient sampling for asymptotic minimum species richness estimators, Ecology'. *Ecology* 90 (2009), 1125–1133.
A. Daly, J. Baetens & B. De Baets, ‘Ecological Diversity: Measuring the Unmeasurable’, *Mathematics* 119 (2018), doi:10.3390/math6070119.
L. Egghe & G. Proot, 'The estimation of the number of lost multi-copy documents: A new type of informetrics theory' Journal of Informetrics 1 (2007), 257-268.
H. Kienhorst, *De handschriften van de Middelnederlandse ridderepiek. Een codicologische beschrijving*. 2 dln. Sub Rosa, 1988.
F. van Oostrom, *Geschiedenis van de Nederlandse literatuur, vanaf het begin tot 1300*. Amsterdam, 2006.