“I Will Survive”: a Computational Study of the Conservation of Printed Text from the Early Modern period in the Netherlands

Arjan van Dalfsen (Utrecht University) @ArjanvD95
Folgert Karsdorp (KNAW Meertens Instituut) @FolgertK
Els Stronks (Utrecht University)

The total amount of historical texts that survived (i.e. “can be consulted in (digital) libraries and archives”) represents only part of the copies that were actually produced at the time. The "lost" copies have become somewhat of a hype in the past decade or so (Bruni & Pettegree, 2016; Egghe & Proot, 2007; Green et al., 2011; Hill, 2018; Kestemont & Karsdorp, 2020; Kestemont et al., 2022; Pettegree & Der Weduwen, 2018; Proot, 2016; Wilkinson, 2009). Perhaps due to the fact that more and more texts were digitized, and research with digitized corpora is now doable, the question arose to what extent the surviving copies are an accurate representation of the original corpus. This question is relevant because the representativeness of the corpus determines the generalizability of research and because patterns of loss can give insight into the dynamics of cultural memory: what texts were so important that they were carefully preserved and what texts were deemed unimportant?

Research into the loss of texts or textual genres is done along two lines that I would label as a descriptive and an inferential one. In descriptive research, one looks at contemporaneous survey lists, such as catalogs, and uses them to estimate what percentage of historical texts have been preserved. An example of this type of research is Pettegree & Der Weduwen (2018). They attempted to estimate the total book production in the seventeenth-century Dutch Republic using for instance catalogs and advertisements. In contrast, research based on the inferential method uses frequencies of texts (how many texts survived in one, two, three, etc. copies) to produce an estimate of the number of texts that survived in zero copies (and thus vanished from our memory). An example of such research is Kestemont et al. (2020). In this research, formulas (“estimators”) from ecology were used to estimate the “unseen” percentage for medieval Dutch chivalric novels. The descriptive method extrapolates from known artifacts, the inferential method from a pattern.

Presented here is inferential research on texts produced in the Dutch Republic, between 1550-1800, in which the ecological estimator CHAO1 (Chao, 1984) is applied to the *Short Title Catalogue, Netherlands* (STCN), the retrospective, historical bibliography of the Netherlands (Bos & Gruys, 2009). Because of the immense size of *STCN* (approximately 600,000 copies, belonging to on about 200,000 editions), this research is restricted to unique editions (the number of copies produced by the printer in one edition). This restriction is an option because the *STCN* is able to distinguish between various editions made of one single text by exploiting a “textual fingerprint” (“Short-Title Catalogue Netherlands (STCN)”, n.d.).

Our research shows that in 2022, in terms of unique editions, the *STCN* covers at most 66% of the number of editions produced at the time. Even taking into account that (due to the theoretical underpinnings of the CHAO1 estimator) this number is an 'at most' estimate, it seems justified to conclude that sufficient texts have been preserved in the *STCN* to label the STCN as a trustworthy source for literary historians when it comes to representativeness. Especially when we also consider that 66% of the total number of editions covers more than 66% of the total number of unique works; after all, many works survived in multiple editions. The coverage ratio calculated with our method is higher than expected based on previous research on the *STCN* (Pettegree & Der Weduwen, 2018).

In addition, sub-corpora in the *STCN* were examined. The survival rate varies substantially among the sub-collections: some sub-corpora are almost completely preserved, while others were almost wiped out. Sub-corpora were based on factors such as language, source language (in the case of translations), format, year of publications, typographic features, subjects, printers, authors and size. In some cases, our research yields results that conflict with existing estimates of text survival. For example, the claim that foreign language texts were better preserved than texts in the vernacular, that larger formats had better chances of survival than smaller ones, and that older texts were more likely to be lost (Harris, 2007) were contradicted by our findings. Other claims, such as the enormous loss rates of books destined for practical use (Harris, 2007; Pettegree, 2016), are confirmed in this study. We find that the strongest effect on survival is seen with the size variable (that is, the paper required to print a text): the larger the size, the more often the text survived.

All in all, our research has yielded a great deal of information about the survival rate of texts printed in the Dutch Republic. We have new estimates for the total of produced texts as well as for various sub-corpora. In general, the news is good: the percentage of texts that survived are generally high enough to assume that the surviving texts form a representative selection of the historic reality. Some sub-corpora are poorly preserved: in this case, it is important for linguists and historians to account for this in their analyses. With regard to the factors causing survival, nothing definitive can be concluded because correlation and causation are still difficult to separate here. Existing literature has suggested that books purchased by wealthy people were best preserved (Proot, 2016). The results provide support for that proposition but do not prove it. To do so, additional research on a causal model for text loss is desirable.

[Number of words: 880 (incl. in-text citations)]

Zenodo Repository
https://zenodo.org/record/6778683#.Y-JUkXbMJD8

References

Bos, J., & Gruys, J.A. (2009). Veertig Jaar STCN 1969-2009. *Jaarboek Voor Nederlandse Boekgeschiedenis*, 16, 9-36. https://www.dbnl.org/tekst/_jaa008200901_01/_jaa008200901_01_0002.php.

Bruni, F., & Pettegree, A. (2016). *Lost Books: Reconstructing the Print World of Pre-Industrial Europe*. Brill Academic Publishers. https://doi.org/10.1163/9789004311824.

Egghe, L., & Proot, J. (2007).The Estimation of the Number of Lost Multi-Copy Documents: A New Type of Informetrics Theory. *Journal of Informetrics*, 1(4), 257-268. https://doi.org/10.1016/j.joi.2007.02.003

Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics 11*(4), 265-270.

Green, J., McIntyre, F., & Needham, P. (2011). The Shape of Incunable Survival and Statistical Estimation of Lost Editions. *The Papers of the Bibliographical Society of America*, 105(2), 141-175. https://doi:10.1086/680773

Harris, N. (2007). La Sopravvivenza Del Libro, Ossia Appunti Per Una Lista Delle Lavandaia. *Ecdotica*, 4, 24-65.

Hill, A. (2018). *Lost Books and Printing in London, 1557-1640: An Analysis of the Stationers' Company Register*. Brill Academic Publishers. https://doi.org/10.1163/9789004349209

Kestemont, M., & Karsdorp, F. (2020). Estimating the Loss of Medieval Literature with an Unseen Species Model from Ecodiversity. *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, 44-55. https://doi.org/10.5281/zenodo.4030681

Kestemont, M., Karsdorp F., de Bruijn, E., Driscoll, M., Kapitan, K., Ó Macháin, P., Sawyer, D., Sleiderink, R., Chao, A. (2022). Forgotten books: the application of unseen species models to the survival of culture. *Science* 375(6582), 765-769. doi:10.1126/science.abl7655

Mathijsen, M. (2010). *Naar De Letter. Handboek Editiewetenschap* (4th ed.). KNAW Press.

Pettegree, A. & Der Weduwen, A. (2018). What was Published in the Seventeenth-Century Dutch Republic? *Livre. Revue Historique, Société Bibliographique De France*, 1-22.
Pettegree, A. (2016). The Legion of the Lost. Recovering the Lost Books of Early Modern Europe. In A. Pettegree & F. Bruni (Eds.), *Lost Books: Reconstructing the Print World of Pre-Industrial Europe* (pp. 1-27). Brill Academic Publishers. http://dx.doi.org/10.1163/9789004311824_002.

Proot, J. Survival Factors of Seventeenth-Century Hand-Press Books Published in the Southern Netherlands: The Importance of Sheet Counts, Sammelbände and the Role of

Institutional Collections. In A. Pettegree & F. Bruni (Eds.), *Lost Books: Reconstructing the Print World of Pre-Industrial Europe* (pp. 160-201). Brill Academic Publishers. http://dx.doi.org/10.1163/9789004311824_009.

Short-Title Catalogue Netherlands (STCN). (n.d.). Retrieved from https://www.kb.nl/over-ons/diensten/stcn.

Wilkinson, A.S. (2009). Lost Books Printed in French before 1601. *The Library* 10(2), 188-205. http://doi:10.1093/library/10.2.188