On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology

Marcel Bollmann
Department of Computer Science
University of Copenhagen
marcel@di.ku.dk

Desmond Elliott
Department of Computer Science
University of Copenhagen
de@di.ku.dk

Abstract

The field of natural language processing is experiencing a period of unprecedented growth, and with it a surge of published papers. This represents an opportunity for us to take stock of how we cite the work of other researchers, and whether this growth comes at the expense of "forgetting" about older literature. In this paper, we address this question through bibliographic analysis. We analyze the age of outgoing citations in papers published at selected ACL venues between 2010 and 2019, finding that there is indeed a tendency for recent papers to cite more recent work, but the rate at which papers older than 15 years are cited has remained relatively stable.

1 Introduction

> "This paper does not cite any literature from before the neural network era."

Scientific progress benefits from researchers "standing on the shoulders of giants" and one way for researchers to recognise those shoulders is by citing articles that influence and inform their work. The nature of citations in NLP publications has previously been analysed with regards to topic areas (Anderson et al., 2012; Gollapalli and Li, 2015; Mariani et al., 2019b), semantic relations (Gábor et al., 2016), gender issues (Vogel and Jurafsky, 2012; Schluter, 2018), the role of sharing software (Wieling et al., 2018), and citation and collaboration networks (Radev et al., 2016; Mariani et al., 2019a). Mohammad (2019) provides the most recent analysis of the ACL Anthology, looking at demographics, topic areas, and research impact via citation analysis.

In this paper, we conduct a corpus analysis of papers published in recent ACL venues to determine whether the community is collectively forgetting about older papers as it experiences a period of rapid growth (see Figure 1). The Association of Computational Linguistics (ACL) is one of the largest publishers of articles in natural language processing research: it maintains the open-access ACL Anthology[1] of articles that date back to the 1960s, offering a rich resource for studying NLP publications. While the aforementioned analyses have mainly focused on *incoming* citations, our work targets *outgoing* citations. We focus on the age of citations in the References section of articles published at ACL venues between 2010 and 2019 (Sec. 2), with a view to studying three questions:

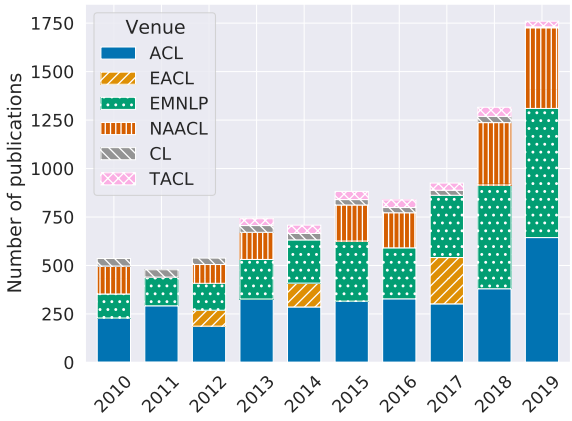

Figure 1: The distribution of the number of articles published between 2010–2019 in the ACL Anthology.

1. Do recently published papers have a tendency to cite more recently published papers, and less older literature?
2. Are older papers being cited less frequently in 2019 than they were in 2010?
3. Is there a difference between publication venues with regard to the age of citations?

We find that the mean age of the papers cited does indeed decrease from 2010–2019, and that this de-

[1] https://www.aclweb.org/anthology/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7819–7827
July 5 - 10, 2020.

crease is statistically significant, with a larger effect size in recent years (Sec. 3.1). We also find that there is no significant difference in the rate at which older papers are cited during this period (Sec. 3.2), and that there are marked differences between the citations in journal articles and conference proceedings (Sec. 3.3). Our findings show that, at a time of rapid growth, an increasing proportion of citations are going to recently published papers, but that researchers still acknowledge that they are standing on the shoulders of their peers.

2 Data

The analysis in this paper is based on a subset of articles from the ACL Anthology. While corpora of NLP publications, including the ACL Anthology, already exist (Bird et al., 2008; Radev et al., 2009; Mariani et al., 2019a), none of them include publications newer than 2015. We compiled our own dataset because we are mostly interested in the papers published in recent years.

The dataset is drawn from ACL venues: conference proceedings from meetings of the ACL, EACL (European Chapter of the ACL), NAACL (North American Chapter of the ACL), and EMNLP (Empirical Methods in NLP) as well as articles from the CL (Computational Linguistics) and TACL (Transactions of the ACL) journals.

Anthology statistics Figure 1 shows the distribution of the articles in the corpus: the number of articles published in these venues steadily increases from 2010–2019. The CL and TACL journals publish articles at a steady rate; the ACL conference fluctuates in size, depending on whether it is co-located with NAACL; and the EACL conference nearly doubles in size each time it takes place. In terms of whether the field is rapidly growing, we note that there was a year-on-year increase of 42% between in 2017–2018 due to the increase in the number of papers published at NAACL and EMNLP, and a 34% increase between 2018–2019.

Extracting citations To extract a list of references from an article, we first extract the text stream from the PDF file via `pdftotext`,[2] then feed it into ParsCit (Councill et al., 2008) to obtain the references.[3] For each reference in this list, we then extract and keep the parsed "date", "author", and "title" entries. For 1.4% of the input files, this pipeline fails to extract any references; spot-checking reveals that many of those are not regular papers (but, e.g., book reviews or front matter), some PDFs have no embedded text, and others silently fail to parse.

Citation age For each publication in our dataset, we want to consider how recently each paper in its reference list was published. We calculate the *age* of a cited paper by subtracting its year of publication from that of the citing paper. We only keep citations in the age range $[0, 50]$ as values outside of this range typically appeared to be parsing errors.[4] As only 0.95% of parsed reference dates fall outside of this range, the effect of excluding potentially valid citations is minimal.

Identifying cited papers We use authors and titles of cited papers in order to identify which individual papers are being cited. We find that these entries are rather noisy in our ParsCit output; therefore, we use a heuristic based on fuzzy string matching to identify citations that are likely to refer to the same paper, despite differences in their author and/or title fields.[5]

Dataset[6] The resulting dataset covers 8,722 papers published within 2010–2019 with a total of 264,957 extracted citations;[7] for conference proceedings, we only include volumes that are marked as containing either *full papers* or *short papers*.[8]

3 Analysis

3.1 Are more recently published papers citing more recently published papers?

Figure 2 shows the distribution of the age of cited articles with respect to the year in which the source article was published; Table 1 gives some complementary statistics. The mean age of a cited paper has steadily decreased since 2013, from 7.69 years to 5.53 years in 2019; the median has dropped from 6 to 3 years in the same period.

[2] https://gitlab.freedesktop.org/poppler/poppler

[3] We note that the ParsCit maintainers recommend a newer iteration of the tool, Neural-ParsCit (Prasad et al., 2018), but we could not easily replicate the same pipeline with it.

[4] For example, ParsCit mistakes the journal number for the year of publication, resulting in a ∼1,900 years old citation.

[5] The full algorithm is described in Appendix A.

[6] Datasets and code are available at: https://github.com/coastalcph/acl-citations

[7] This includes papers that were published on the ACL Anthology before November 6, 2019.

[8] In particular, this excludes papers from system demonstration, student research workshop, and industry tracks.

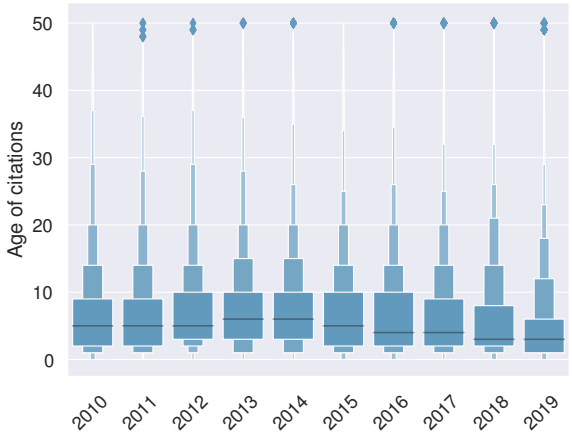

Figure 2: Letter-value plot (Hofmann et al., 2017) showing the distribution of citation ages in the corpus, grouped by year of publication. The solid black lines denote the median, boxes correspond to quantiles.

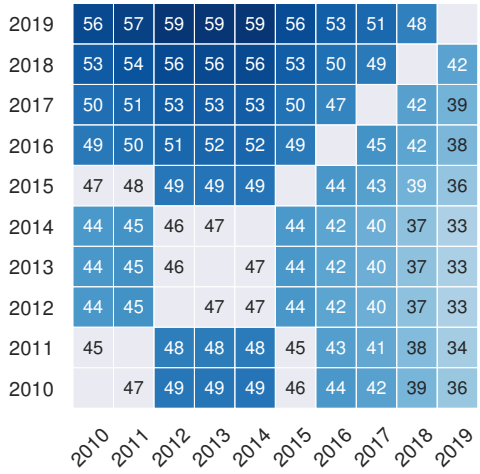

Figure 3: Common language effect size (CLES) scores for the distribution of citation age (cf. Sec. 3.1 for interpretation); greyed-out cells indicate pairs where the difference in distribution was not statistically significant.

Significance and effect size To determine if the distribution of citation ages significantly differs between years, we perform Mann-Whitney U tests with $p < 0.005$ and Bonferroni correction on each pair of years. We calculate rank-biserial correlation scores to determine the effect size of these differences and convert them into common language effect size (CLES; McGraw and Wong, 1992) for easier interpretability.[9] Results are shown in Figure 3: numbers correspond to (rounded) CLES values and can be interpreted as the probability that a randomly drawn citation from the *column* year will be older than a randomly drawn citation from the *row* year. For example, if we were to randomly draw a citation from a paper published in 2012 and one from a paper published in 2019, the former citation has a 59% probability of being strictly older than the latter (row "2019", column "2012"). Greyed-out cells were *not* statistically significantly different according to the Mann-Whitney U test.

The CLES scores show that a randomly drawn citation from more recent years (e.g. 2017–2019) has a significantly lower probability of being older than a randomly drawn citation from earlier years (e.g. 2010–2014). This can be seen by inspecting the columns and rows in the bottom right of Figure 3.

3.2 Are older papers cited less frequently in more recently published papers?

While the previous section showed a downwards trend for average citation age in more recent publications, this does not imply that older papers are cited less frequently in *absolute* terms. Indeed, as there are more publications available to cite from recent years, it seems natural that they would constitute a larger *relative* share of cited papers, but this does not necessarily need to come at the cost of citing older papers less frequently.

		Citation Age		
Year	**Count**	Median	Mean	SE
2010	12,919	5	7.27	.068
2011	12,662	5	7.38	.068
2012	14,679	5	7.63	.063
2013	21,363	6	7.69	.052
2014	21,208	6	7.66	.051
2015	25,616	5	7.21	.046
2016	26,465	4	7.00	.047
2017	30,511	4	6.69	.043
2018	42,962	3	6.26	.036
2019	56,572	3	5.53	.029

Table 1: Number of citations for each year of publication, along with median age, mean age, and standard error (SE) of the mean.

Figure 4 visualizes the average number of citations per paper, broken down by the age of the citation. We observe that this number steadily increases between 2010 and 2019, showing that publications in 2019 do indeed cite more papers than publications in 2010, on average. We also see that this increase is mostly due to citations of papers between 0 and 3 years old, while papers that were published 15 or more years ago are still cited at approximately the same rate now as in 2010.

[9] If r is the rank-biserial correlation coefficient, CLES is defined as $\frac{r+1}{2}$.

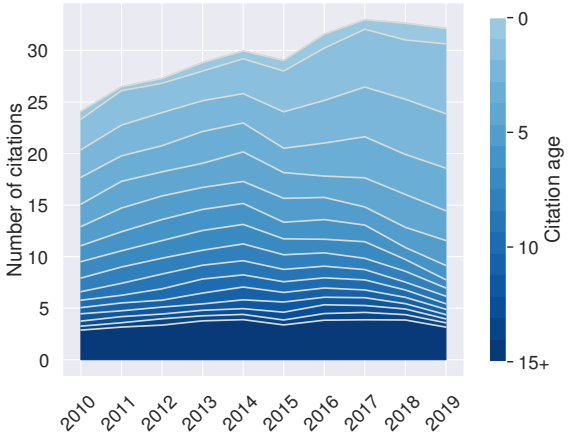

Figure 4: Average number of citations per paper with a given age. Bottom (darkest) area includes all citations of age 15 or older; each area above that represents citations of the next lower age.

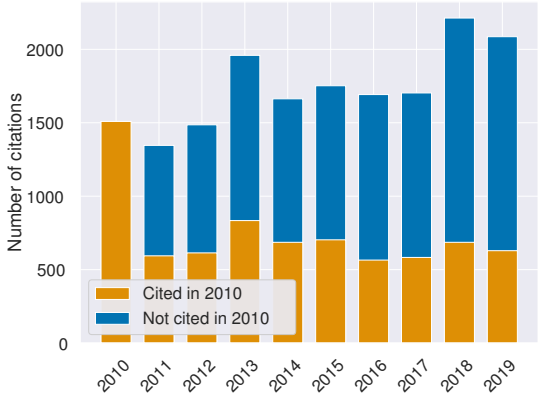

Figure 5: Average number of citations per papers with age 15 or older, distinguished by whether or not they (already) have been cited in 2010.

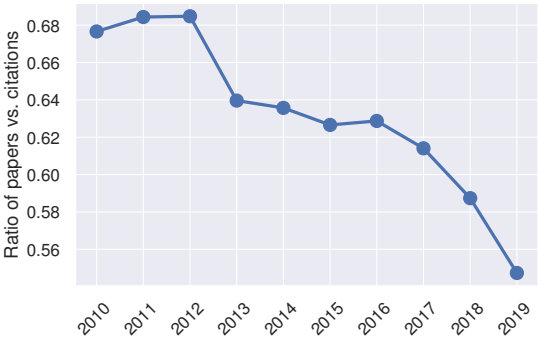

Figure 6: Ratio of unique citations (i.e., papers) and total citations of age 15 or older.

Tracking citations to individual papers While the citation rate for "old" papers has not changed, the distribution of papers being cited may have. To investigate this, we now also consider the author and title fields of citations to track *which* papers are being cited. This way, we can analyze e.g. to what extent "old" papers cited in 2010 overlap with those cited in 2019. Figure 5 shows the average number of citations to papers published 15 or more years ago—corresponding to the bottom area of Fig. 4—and additionally indicates which share of these papers have already been cited in 2010. We can see that in all the other years, more than half of these "old" citations are to papers that were not cited in 2010.

Table 2 shows the most frequently cited "old" papers in 2019, additionally indicating in which year we can find the earliest citation to this paper in our dataset. Perhaps unsurprisingly, the most cited papers describe very broadly applicable resources or methods. Furthermore, two of these papers—introducing the bidirectional RNN and the LSTM, respectively—have only gathered citations from 2014 onwards, while another classic reinforcement learning paper was not cited before 2016. This suggests that in recent years, a substantial part of older citations is made up of deep learning papers that have not yet been (widely) cited in 2010.

Ratio of papers to citations Figure 6 looks at the *ratio* of unique "old" papers being cited compared to the total number of citations. We observe that this ratio has steadily decreased since 2013, indicating that the stable number of citations goes to a continuously decreasing pool of papers. In other words, there is a reduction in the *variety* of older papers being cited.

3.3 Do publication venues differ in how frequently older papers are cited?

Journals invite submissions that are more substantial than conference papers; it is conceivable that this is reflected in the papers they cite. Figure 7 takes a closer look at citations 15 years or older by venue of publication. The four conference venues in our dataset behave very similarly, showing around 2–4 "old" citations on average. For CL papers, on the other hand, this figure is considerably larger (up to 17 such citations on average in 2017). TACL papers also show a trend towards more older citations, but not as strong as for CL. Overall, there is a clear difference in the average number of older citations in journal articles compared to conference proceedings.

Citations	First cited	Paper
250	2010	Papineni et al. (2002). BLEU: a method for automatic evaluation of machine translation.
117	2010	Lin (2004). ROUGE: A package for automatic evaluation of summaries.
91	**2014**	Hochreiter & Schmidhuber (1997). Long short-term memory.
83	**2016**	Williams (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning.
62	2010	Lafferty et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
60	2010	Marcus et al. (1993). Building a large annotated corpus of English: The Penn Treebank.
53	2010	Miller (1995). WordNet: a lexical database for English.
47	2010	Blei et al. (2003). Latent dirichlet allocation.
40	**2014**	Schuster & Paliwal (1997). Bidirectional recurrent neural networks.
39	2010	Hu & Liu (2004). Mining and summarizing customer reviews.

Table 2: The most frequently cited papers in 2019 with citation age 15 or older (i.e., published before 2005). "First cited" is the year of the earliest extracted citation to this paper *in our dataset*.

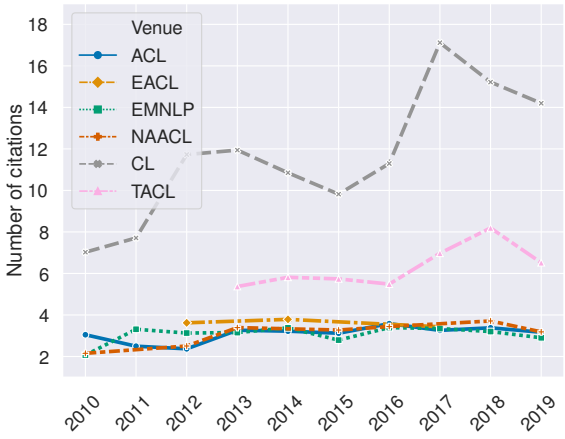

Figure 7: Average number of citations per paper that are 15 years or older, by venue of publication.

4 Conclusions

We presented an analysis of citations in publications from major ACL venues between 2010 and 2019, focusing on the distribution of the age of cited papers. We found that recently published papers (0–3 years old) are cited significantly more often in publications from recent years (ca. 2015–2019), while papers published 15 or more years ago are being cited at a stable rate. There is also a marked difference between journal and conference publications in the distribution of citation age: journal articles feature more citations to older papers.

These findings could be due to the increasing difficulty of keeping up with the literature, given that many more papers are being published now, in addition to the deluge of papers that appear on preprint servers. Some areas of NLP research did also not exist 15 years ago, e.g. social media analysis, potentially making it challenging to cite older related work. Finally, since several influential neural network papers have been published in the 1990s (cf. Tab. 2), a mostly quantitative analysis is limited in its ability to determine, e.g., to what extent we still engage with older literature outside of this domain.

A potential confound in our analysis is that some proceedings imposed a page limit for references; e.g., the ACL conference gave unlimited space for references in 2010, 2012, and from 2016 onwards, but imposed a page limit in 2011 and 2013–2015. We can still observe an increase in the average number of citations per paper during this latter period, so it seems unlikely that this had an effect. In addition, our analysis is limited to studying the age of the papers cited in the ACL Anthology – it does not make any claims about the complex network effects involved in researchers from particular institutions, countries, or sub-fields, and it does not study other venues that also publish NLP papers.

Future work includes a deeper qualitative analysis of *which (type of)* papers are being cited; a more fine-grained analysis of different research topics in NLP to determine whether changes are more prevalent within certain areas than others; or extending the analysis to a larger set of the papers in the ACL Anthology.

Acknowledgments

We would like to thank the reviewers for their helpful comments and suggestions for further analyses. Marcel Bollmann was funded from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 845995.

References

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 2012. Towards a computational history of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21, Jeju Island, Korea. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Isaac Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: an open-source CRF reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 661–667, Marrakech, Morocco. European Language Resources Association (ELRA).

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3694–3701, Portorož, Slovenia. European Language Resources Association (ELRA).

Sujatha Das Gollapalli and Xiaoli Li. 2015. EMNLP versus ACL: Analyzing NLP research over time. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2002–2006, Lisbon, Portugal. Association for Computational Linguistics.

Heike Hofmann, Hadley Wickham, and Karen Kafadar. 2017. Letter-value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, 26(3):469–477.

Joseph Mariani, Gil Francopoulo, and Patrick Paroubek. 2019a. The NLP4NLP corpus (I): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3:36.

Joseph Mariani, Gil Francopoulo, Patrick Paroubek, and Frédéric Vernier. 2019b. The NLP4NLP corpus (II): 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3:37.

Kenneth O. McGraw and S. P. Wong. 1992. A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

Saif M. Mohammad. 2019. The state of NLP literature: A diachronic analysis of the ACL Anthology. *arXiv preprint arXiv:1911.03562*.

Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018. Neural ParsCit: a deep learning-based reference string parser. *International Journal on Digital Libraries*, 19(4):323–337.

Dragomir R. Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2016. A bibliometric and network analysis of the field of computational linguistics. *Journal of the Association for Information Science and Technology*, 67(3):683–706.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology Network Corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore. Association for Computational Linguistics.

Natalie Schluter. 2018. The glass ceiling in NLP. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2793–2798, Brussels, Belgium. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

A Fuzzy paper matching

In Section 3.2, we track citations to individual papers, which requires identifying authors and titles of cited papers in addition to their year. Since this information is rather noisy in the output we obtain from ParsCit, we employ a simple matching algorithm. This algorithm heuristically matches citations with non-identical author and/or title fields which are likely to refer to the same paper.

Concretely, we first preprocess the author and title fields as follows:

1. We convert strings to a pure ASCII representation.[10]
2. We cut off the title field after a *dot–space* (`. `) sequence, as we found this to almost always indicate the start of the journal/proceedings/booktitle field (which was incorrectly interpreted as part of the title by ParsCit).

We then treat two citations as referring to the same paper if all of the following criteria hold:

1. Their year of publication is identical.
2. They have the same number of authors.
3. All author last names can be fuzzy-matched.
4. All author first names can be fuzzy-matched *or* they start with the same character.[11]
5. Their titles can be fuzzy-matched.

Two strings can be *fuzzy-matched* if their distance ratio[12] is $\leq 95\%$.

Quality of paper matching We found the described approach to work reasonably well on our citation data, though it unfortunately still results in many false negatives (i.e., papers that should have been matched but were not). Common problems include:

- Papers that are cited with inconsistent author lists; e.g., the paper that introduced the Penn Treebank is cited as *"Marcus, Santorini, Marcinkiewicz"*, *"Marcus, Marcinkiewicz, Santorini"*, or *"Marcus & Marcinkiewicz"*.
- Papers with both pre-print and peer-reviewed versions that were not published in the same year.
- Parsing or text extraction errors.

[10] We achieve this by using `https://github.com/un33k/python-slugify`.

[11] The motivation here is that some citation styles use full first names, while others only give initials.

[12] As implemented by `https://github.com/seatgeek/fuzzywuzzy`.

B Supplementary figures

B.1 Oldest citation per paper

Figure 8 shows the distribution of the *oldest citation* per paper in our dataset. This is motivated by the idea that while the *average* number of "old" citations per paper is stable (cf. Sec. 3.2), they might be distributed in an unbalanced way. In other words, there might be a subset of publications that does not cite *any* "older" work. Figure 8 shows that this is not really the case: the majority of papers in our dataset include a citation of age 15 or older. There are a few outliers, however: there are 15 papers in total which, according to our processing pipeline (cf. Sec. 2), do not include any citation older than 3 years. We manually check their original PDFs and find that one of these is a book review, three are extraction errors, and 11 actually do not contain any citation older than 3 years.

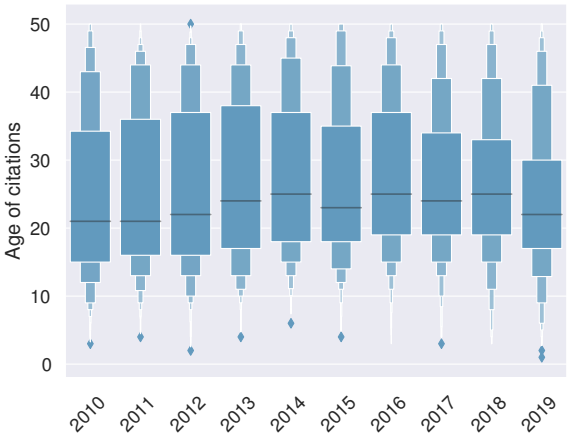

Figure 8: Letter-value plot (Hofmann et al., 2017) considering only the *oldest citation per paper* among all papers published in a given year. The solid black lines denote the median, boxes correspond to quantiles.

B.2 Extended versions of previous figures

Figure 9 shows the distribution of citation ages, analogous to Figure 2, but separately for each publication venue.

Figure 10 shows the average number of citations per paper, analogous to Figure 7, but for a larger number of citation ages.

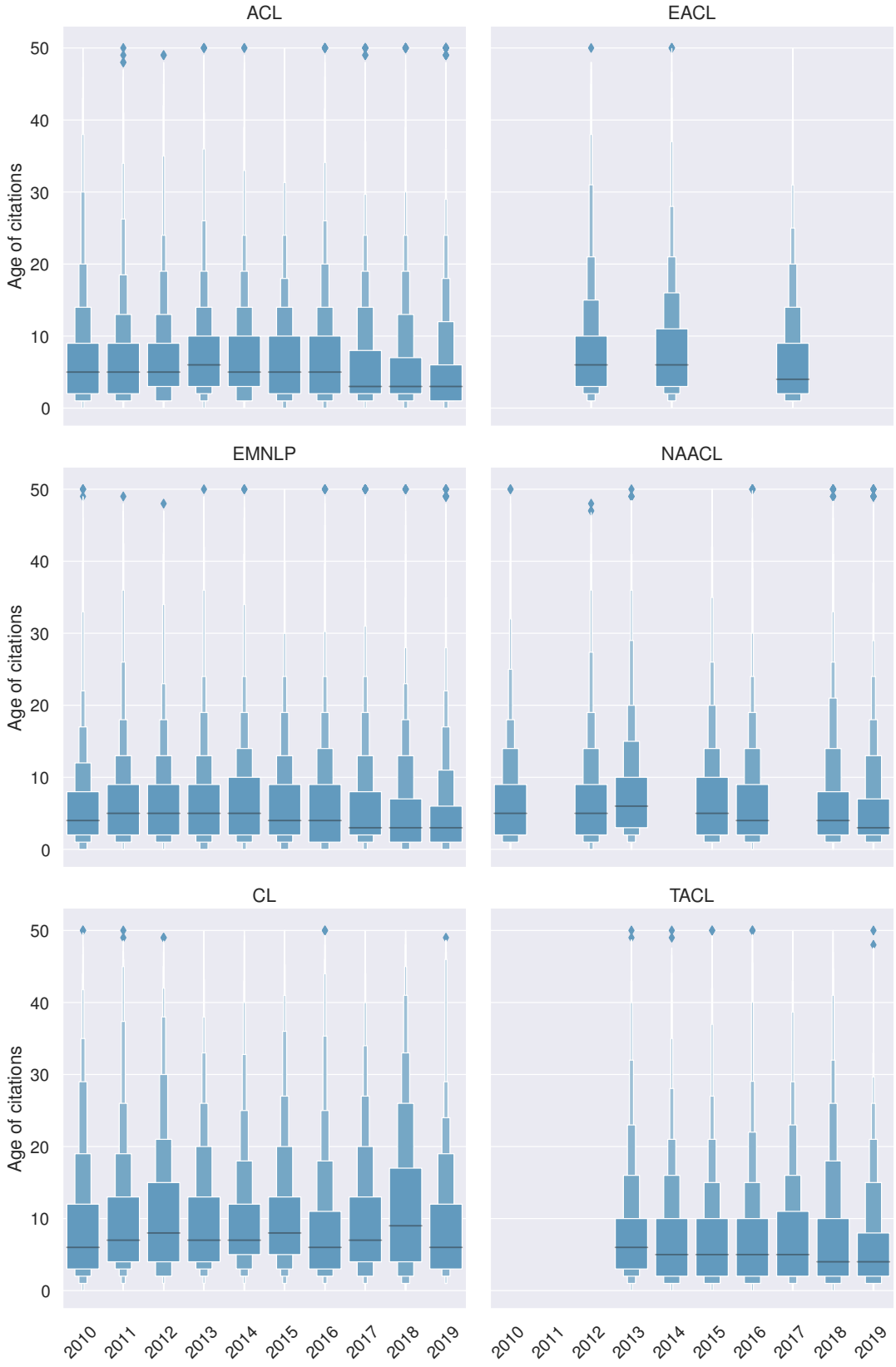

Figure 9: Letter-value plot (Hofmann et al., 2017) showing the distribution of citation ages by publication venue, grouped by year of publication. The solid black lines denote the median, boxes correspond to quantiles.

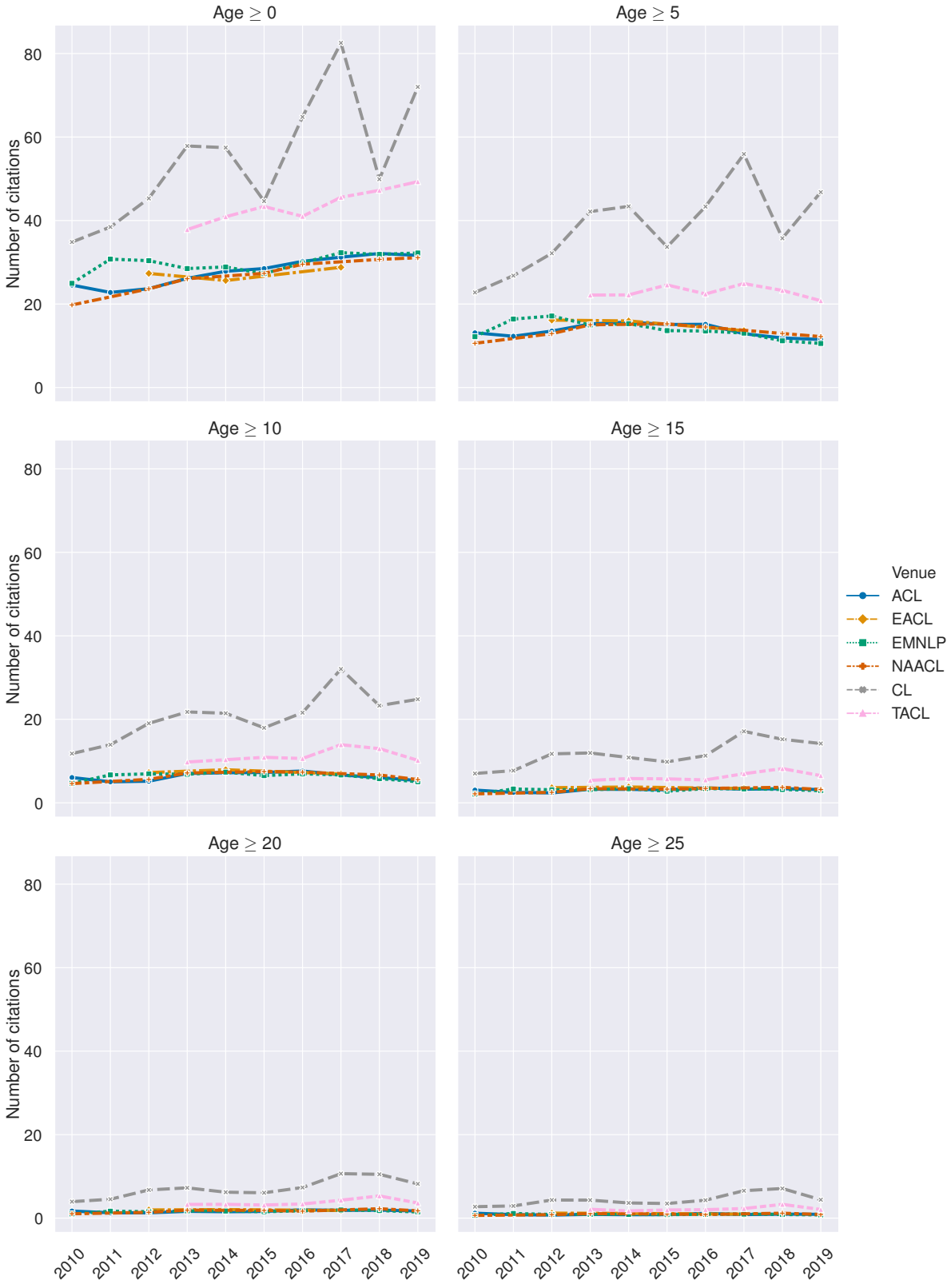

Figure 10: Average number of citations per paper, separately by venue of publication, for a number of different citation ages.