On the approachability of top researchers in computational linguistics

Anonymous ACL submission

Abstract

Computational linguistics and natural language processing research have been in a surge of popularity given the advent of instruction-tuned large language models that nowadays are almost a commodity for laypeople, propelling researchers to unprecedented fame. However, this increased visibility can have an unintended consequence: making it more challenging for early-career or lesser-known researchers to connect with leading experts. This study investigates the approachability of top computational linguistics researchers, examining their willingness to engage with junior researchers. We compiled a list of the top 100 most cited researchers in computational linguistics by analyzing publication data from the ACL anthology and citation counts from OpenAlex. We then conducted a manual search to gather their contact information, including personal websites, email addresses, and social media handles. Next, we sent a request for advice to each researcher with a public email, posing as a PhD student seeking advice regarding the topic of their latest publication. Our results show that 75 % of top researchers maintain a personal website, 82 % have a public email, and only 16 % responded to our email request. We also propose a categorization system to assess the researchers' level of approachability. This study highlights the need to improve the reachability of influential researchers, which could facilitate greater collaboration and knowledge exchange between experienced and early-career researchers, ultimately enriching the next generation of computational linguistics researchers.

1 Introduction

The field of computational linguistics, encompassing natural language processing and human language technologies, has experienced a remarkable surge in popularity with the emergence of large language models (LLMs) capable of instruction-following (Touvron et al., 2023a,b). As LLMs become increasingly accessible to the general public, they fuel a growing demand for high-quality research in this area, propelling the field forward at an unprecedented pace. For example, since October 2023, the Rolling Review of the Association of Computational Linguistics (ACL), a review system for all ACL venues, has been experiencing an unprecedented growth in submissions per cycle (ACL, 2024).

The exceptional significance of computational linguistics research, combined with the influence of social media, has created a perception of unprecedented accessibility to leading researchers (Siebert et al., 2023; Zhang and Lu, 2022). As a result, researchers are increasingly becoming celebrated figures, transcending their academic credentials to attain fame beyond their published work.

The increased visibility of top researchers in the digital sphere has led to a potential tradeoff: their newfound public presence may come at the cost of reduced approachability. Overwhelmed by daily social media and email messages, they may struggle to respond to even the most well-intentioned inquiries. Some have explicitly stated that they do not answer unsolicited messages, effectively establishing a barrier to entry for those seeking to connect with them.

The inaccessibility of prominent researchers can disproportionately impact the career trajectories of young scholars from underrepresented regions, such as low-income countries and the Global South, who may rely heavily on digital communication channels to connect with leading experts. As networking plays a crucial role in shaping the success of early-career researchers (Mazurek et al., 2020; Wolff and Moser, 2009), the unavailability of influential researchers for informal discussions or guidance can exacerbate existing disparities in access to resources, opportunities, and knowledge, ultimately perpetuating the marginalization of already underrepresented groups.

Related work Research has shown that the volume of emails received by faculty members correlates positively with their stress levels, which can, in turn, affect their approachability and responsiveness (Jerejian et al., 2013). An audit study found that faculty members were significantly more responsive to PhD applicants who were Caucasian males, whereas applicants from other gender and racial backgrounds received fewer and slower responses (Milkman et al., 2012a). This disparity in response rates may exacerbate existing inequalities in research opportunities, perpetuating the marginalization of underrepresented groups. Furthermore, another study revealed that male Caucasians receive more and faster responses from faculties when seeking to schedule meetings (Milkman et al., 2012b), underscoring the need to address these biases and promote a more inclusive academic environment.

Outline In this study, we investigate the approachability of top computational linguistics researchers by collecting their publicly available contact information, including personal websites, email addresses, and social media handles. We then sent a standardized email to each researcher who made their email publicly available and measured the response rate. Building on these findings, we propose a categorization system to assess the level of approachability, providing a framework for evaluating their willingness to engage with junior researchers.

2 Data & methods

We queried the ACL anthology database (Gildea et al., 2018), a comprehensive repository of computational linguistics and natural language processing literature, using the official Python library[1] to retrieve all publications' metadata. We then extracted the digital object identifiers (DOIs) associated with each ACL anthology publication, whenever available. Subsequently, we utilized the web service of OpenAlex, a vast bibliographic catalogue of scientific papers, authors, and institutions (Priem et al., 2022), to retrieve the author information and citation counts for each publication.

We aggregated the citation counts for each author across all their publications in the ACL anthology, generating a comprehensive list of authors paired with their total citation counts. From this list, we identified the top 100 most cited authors we selected as our study's focus.

To gather contact information for each of the top 100 authors, we conducted a manual search engine query using their names and retrieved their personal websites. When available, we extracted their email address and social media handle from their website. In cases where the email address was not available on their website, we turned to OpenAlex and examined their publications' PDFs from the last two years to obtain their email address. Similarly, if their social media handle was not listed on their website, we employed a search engine to locate it.

We then sent a standardized email from a university email service to each of the 100 authors (the email template is available in Appendix A), posing as a PhD student seeking advice regarding their most recent publication. In the email, we requested the opportunity to discuss their work in person at one of the upcoming computational linguistics conferences, allowing us to gauge their willingness to engage with junior researchers.

Finally, we describe the findings regarding how many researchers had their website, email, and social media handles available to the public and the response rate to the sent email.

Finally, we categorize the authors in one of the following proposed categories regarding their approachability:

Highly Approachable Researchers who responded to our email inquiry within seven days demonstrated high responsiveness and willingness to engage.

Moderately Approachable Researchers who publicly shared their email addresses on their personal websites, making them easily accessible to others.

Somewhat Approachable Researchers whose email addresses were available in publications from the last two years, albeit requiring some effort to obtain.

Less Approachable Researchers who did not make their email addresses publicly available through any channel, making it more challenging for others to contact them.

[1] https://github.com/acl-org/acl-anthology/tree/master/python

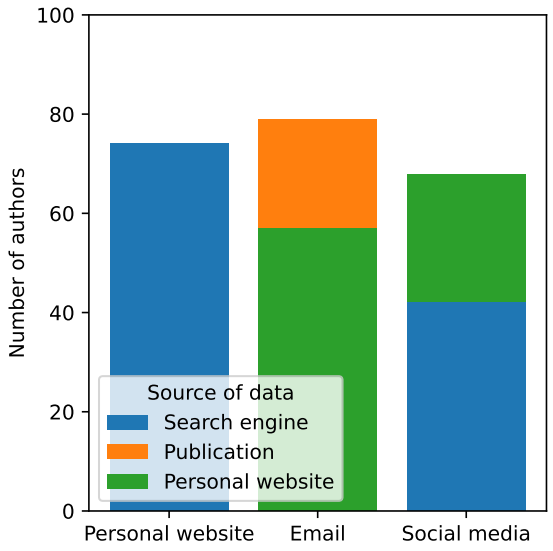

Figure 1: Number of authors with a personal website, email, or social media handle publicly available along with the source of each data.

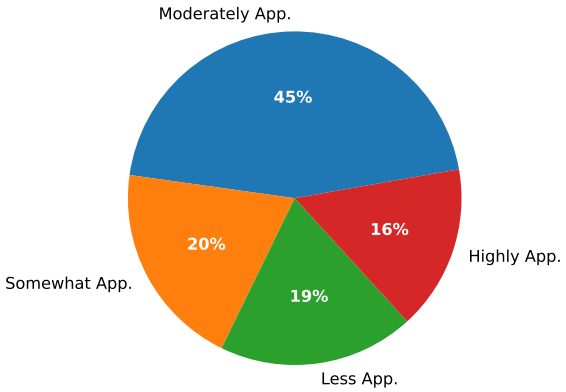

Figure 2: Proportion for each approachability category. App.: Approachable.

3 Results & discussion

Out of 97,345 publications listed in the ACL anthology, 39,544 had valid DOIs, for which we successfully retrieved data from OpenAlex. The absence of valid DOIs for more than half of the database is because the ACL anthology only started to assign DOIs in 2014.

We identified 51,145 unique authors who published papers in the ACL anthology, with a median citation count of 6 per author. Approximately 10,000 authors received no citations at all. To be in the top 1 % of the most cited authors, an individual needed at least 1,484 citations. The highest-cited author gathered 41,108 citations, while the 100th author on the list had 5,014 citations. The citation distribution reveals that most authors received six or fewer citations, with a notable long tail of outliers with thousands of citations; this power law-like distribution has been described widely in the literature (Brzezinski, 2015; Gupta et al., 2005).

All authors are affiliated with institutions in developed countries of the Global North, except for one author affiliated with an institution in Australia (also considered a developed country) and another author from Algeria, the sole representative from a developing country in the Global South among the listed affiliations. This discovery reinforces that high-quality research is disproportionately concentrated in developed countries, perpetuating the power imbalance in global knowledge. The fact that most authors come from these nations suggests they possess the necessary resources, infrastructure, and expertise to produce cutting-edge research, while many developing countries may face significant barriers. Moreover, a striking gender imbalance is evident, with women comprising a mere 12 % of the researchers, underscoring the persistent underrepresentation of females in computer science-related fields. The disparities in computer science, including those related to geographical distribution and gender representation, have been extensively documented in many existing studies (Lietz et al., 2024; Kallia and Cutts, 2021).

Figure 1 illustrates the contact availability of the researchers, showcasing the distribution of publicly available contact information and the sources from which the data were obtained. Most of the researchers' contact information is readily available from personal websites and search engines, and this widespread accessibility of contact details initially suggests that the authors are open to communication and collaboration.

Figure 3 displays the actual response rate to the sent emails, providing a breakdown of the delivery and response rates. The figure shows that only 16 % of researchers responded to our emails, while only four emails were undeliverable due to nonexistent mailboxes. Notably, university email providers dominated the list, and we received a significantly higher response rate from that category compared to company-provided emails.

Figure 3: Diagram of the email deliverability and response rate.

The low response rate observed in this study suggests that researchers may be hesitant to engage with external requests or not have had the time to respond. This could be attributed to various factors, including the increasingly high volume of emails they receive, their busy schedules, or a general lack of interest in collaborating with external parties. The finding that university-affiliated email providers yielded a higher response rate than company-provided emails may indicate that researchers within academic institutions are more receptive to collaborations or have more flexible schedules, allowing them to respond to external requests. Conversely, company-affiliated researchers may be bound by stricter policies or have limited time.

One potential strategy for early-career researchers is initiating connections with academics rather than industry researchers, which may foster more productive and meaningful initial networking experiences.

Figure 2 illustrates the distribution of our proposed approachability categorization, revealing that most authors fall into the Moderately Approachable category. This classification indicates that these researchers have made their email addresses publicly available on their personal websites but failed to respond to our initial contact within a week. Furthermore, an analysis of the authors classified as Highly Approachable, who responded to our email within a week, reveals that most of these researchers have made their email addresses publicly available on their personal websites.

Early career researchers may also consider targeting authors who have proactively shared their email addresses on their personal websites, thus facilitating initial contact.

4 Conclusion

Very few of the top researchers in computational linguistics replied to the request for advice from a PhD student; this highlights a reality where early-career researchers often face significant barriers when attempting to engage with experts in their field. This phenomenon may be even more pronounced for specific demographic groups, such as underrepresented racial and gender minorities, as suggested by the literature. The implications are far-reaching, hinting at a potential lack of access to valuable guidance, mentorship, and collaboration opportunities that can hinder the growth and development of early-career researchers from underrepresented communities.

Our study demonstrated the importance of targeting the right audience for networking opportunities. Specifically, initiating connections with academics rather than industry researchers may foster more productive experiences, as they are more likely to be invested in knowledge sharing and collaboration. Furthermore, our study suggests that targeting authors who have proactively shared their email addresses on their personal websites may facilitate initial contact and increase response rates. By adopting these strategies, early-career researchers can increase their chances of establishing meaningful connections with established experts, ultimately enriching their research experiences and outcomes.

5 Limitations

The sample size is limited to the top 100 most cited researchers in computational linguistics, which

may not be representative of the entire field. Additionally, our evaluation method relies solely on a standardized email, which may not capture the full range of scenarios in which researchers may engage with junior colleagues.

The study also did not investigate the underlying reasons for researchers' approachability, and the categorization system may not be exhaustive. Further research is needed to understand these factors and inform strategies for improving researcher approachability.

References

ACL. 2024. ARR statistics. https://stats.aclrollingreview.org/. [Accessed 06-07-2024].

Michal Brzezinski. 2015. Power laws in citation distributions: evidence from scopus. *Scientometrics*, 103(1):213–228.

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.

Hari M. Gupta, José R. Campanha, and Rosana A. G. Pesce. 2005. Power-law distributions for the citation index of scientific publications and scientists. *Brazilian Journal of Physics*, 35(4a):981–986.

Ailsa C.M. Jerejian, Carly Reid, and Clare S. Rees. 2013. The contribution of email volume, email management strategies and propensity to worry in predicting email stress among academics. *Computers in Human Behavior*, 29(3):991–996.

Maria Kallia and Quintin Cutts. 2021. Re-examining inequalities in computer science participation from a bourdieusian sociological perspective. In *Proceedings of the 17th ACM Conference on International Computing Education Research*, ICER 2021. ACM.

Haiko Lietz, Mohsen Jadidi, Daniel Kostic, Milena Tsvetkova, and Claudia Wagner. 2024. Individual and gender inequality in computer science: A career study of cohorts from 1970 to 2000. *Quantitative Science Studies*, 5(1):128–152.

Grzegorz Mazurek, Anna Gorska, Pawel Korzynski, and Susana Silva. 2020. Social networking sites and researcher's success. *Journal of Computer Information Systems*, 62(2):259–266.

Katherine L. Milkman, Modupe Akinola, and Dolly Chugh. 2012a. Heterogeneity in discrimination?: A field experiment. *SSRN Electronic Journal*.

Katherine L. Milkman, Modupe Akinola, and Dolly Chugh. 2012b. Temporal distance and discrimination: An audit study in academia. *Psychological Science*, 23(7):710–717.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *Preprint*, arXiv:2205.01833.

Maximilian Siebert, Leonardo Maria Siena, and John P.A. Ioannidis. 2023. Twitter and mastodon presence of highly-cited scientists.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Hans-Georg Wolff and Klaus Moser. 2009. Effects of networking on career success: A longitudinal study. *Journal of Applied Psychology*, 94(1):196–206.

Annie Li Zhang and Hang Lu. 2022. Behind the lab coat: How scientists' self-disclosure on twitter influences source perceptions, tweet engagement, and scientific attitudes through social presence. *New Media amp; Society*, page 146144482211416.

A Email template

Template for the email sent to the researchers:

Subject: Kind advice for a new research direction on a topic you recently published on

Dear Dr. `<researcher_name>`,

I hope this email finds you well. My name is [REDACTED FOR ANONIMITY], and I am a PhD student in Computer Science at [REDACTED FOR ANONIMITY]. I am reaching out because I am extremely interested in exploring a new research direction, and your work has

been a significant inspiration to me. In particular, I was interested in your recent publication, "<last_publication_title>", and I believe your expertise would be invaluable in helping me get started in this area.

I would be honoured to discuss my ideas with you in person and learn from your experiences. I noticed that ACL 2024 and EMNLP 2024 are approaching, and I wanted to know if you plan to attend either of these conferences. If so, I would greatly appreciate the opportunity to meet with you for a brief chat.

I understand that your time is valuable, and I appreciate your consideration. I look forward to discussing my research interests with you.

Best regards,

[REDACTED FOR ANONIMITY]

PhD student

[REDACTED FOR ANONIMITY]