Towards better interdisciplinary science: Learnings from COLING 2018

Leon Derczynski, IT University of Copenhagen, Denmark. ld@itu.dk
Emily M. Bender, University of Washington, USA. ebender@uw.edu

Technical report TR-2021-208
ISBN: 978-87-7949-375-9

IT UNIVERSITY OF COPENHAGEN

Abstract

In our role as program committee co-chairs for COLING 2018, we tried out several innovations in the process of creating the conference program, with the following goals: (1) creating a program of high quality papers which represent diverse approaches to and applications of computational linguistics written and presented by researchers from throughout our international community; (2) facilitating thoughtful reviewing which is both informative to area chairs (and PC co-chairs) and helpful to authors; and (3) ensuring that the results published at COLING 2018 are as reproducible as possible. This short paper outlines the innovations and reflects on the ways in which they helped (or didn't) to achieve those goals.

Towards better interdisciplinary science: Learnings from COLING 2018

1 Introduction

The way computational linguistics research is reviewed and published needs to constantly adapt to the field, but practices remained largely static for much of the 2000s and 2010s. This paper describes innovations we introduced and practices we adopted in our process for constructing the program for COLING 2018 [2], which took place in Santa Fe, NM, USA in August 2018, building on previous computational linguistics conferences.

COLING is the oldest computational linguistics conference, running roughly biannually since 1965. It is overseen by the International Committee on Computational Linguistics (ICCL)[1] which maintains a culture of cheerful international collaboration which eschews bureaucracy or regulations. One consequence of this is that COLING PC chairs have great leeway in how to approach the task. We used this opportunity to promote certain best practices in scholarship and counteract pressures from being in a field where expectations of frequent and rapid publication are normalized. We hope that this documentation of our processes may inspire future program chairs to build on these ideas, and will help demystify some aspects of publishing in our field for earlier-career researchers.

We sought to attract and publish a collection of papers that reflected diverse and interdisciplinary perspectives on computational linguistics, that included contributions from all parts of our international community, and that maintained high standards of reproducibility. At the same time, we undertook this role in the spirit of community service and endeavored to make our work both transparent to the community and responsive to community input. The primary vehicle we used for community engagement was our PC chairs blog,[2] in which we took inspi-

[1] https://ufal.mff.cuni.cz/iccl

[2] http://coling2018.org/category/pc-blog/; We link to several blog posts in the footnotes in this paper. Where convenient, these are presented as hypertext links rather than URLs.

ration from the blog produced by Min-Yen Kan and Regina Barzilay for ACL 2017.[3] Wherever possible, we announced program innovations with sufficient lead time to incorporate feedback and we are very grateful to colleagues who took the time to engage with our blog posts.

In this paper we highlight COLING 2018 design decisions and seek to illuminate the extent to which they were effective in meeting our goals. In §2, we describe the strategy of using paper types (and associated review forms) to broaden program content. In addition to diverse program content, COLING 2018 emphasized research quality, as described in §3. High quality research is carried out all around the world and on many different languages, but papers written about languages other than English and/or by people with less experience with academic English frequently face higher barriers to publication. Section 4 describes our strategies for addressing this. In §§5–7 we lay out our reviewing process, addressing how we maintained anonymity from start to finish, how we sought to improve consistency and fairness of reviewing, and how we used reviewer input to determine acceptance. Finally, in §8, we present our process for determining best paper awards, designed to recognize a broad variety of excellence in our field.

2 Making space for more diverse contribution types

Computational linguistics is an interdisciplinary field and of all our conferences, COLING is perhaps the best series of events for bringing people from the different relevant disciplines together and providing a place for them to interact. But for any given iteration of the conference, this promise of interdisciplinarity can only be achieved if the conference is able to attract research from different perspectives. From our prior experience as both authors and reviewers, we had the sense that the one-size-fits-all review forms typical in current NLP conferences made it more difficult to get papers outside of the current dominant type accepted. Therefore, and taking inspiration from the work of Sandra Carberry and Stephen Clark as PC chairs of ACL 2010, we developed a series of six paper types, each with its own associated review form.

In developing the paper types, we had several (sometimes conflicting) design goals, including: (1) defining paper types and associated

[3] https://acl2017.wordpress.com/

review forms that were specific enough to both attract and fairly evaluate diverse contributions; (2) providing a broad enough range of paper types that everything typically considered computational linguistics had a home; (3) keeping the number of different paper types small enough to be navigable; and (4) making sure the community at large was aware of the paper types and understood the intent of each.

In order to meet these design goals, we leaned on our PC blog as a means of community engagement. We published a post with an initial set of five paper types and associated review forms on August 17, 2017.[4] We proposed five initial paper types:[5] computer-aided linguistic analysis paper, NLP engineering experiment paper, reproduction paper, resource paper, and position paper. Based on the feedback, we refined each type's review form questions, and added a sixth paper type (survey paper). Importantly, paper types are not tracks: we used tracks driven by paper topic (described in §6) to assign reviewers to papers; we used paper types to assign review *forms* to papers.

Most of the feedback to the paper types idea was extremely positive,[6] but we encountered some concerns that we want to surface here. First, some authors noted that it was hard to choose between the paper types, with their papers seeming to span more than one. While we understand this difficulty, we think that having multiple review forms to choose from is an improvement over having to use just one that might favour some types of work over others. One form is unlikely to fit all kinds of work. Second, it was suggested by Ron Artstein that authors might use the review forms to shape their papers. We see this as a feature of the approach, especially in the case of authors new to the field. Finally, there were suggestions for additional paper types, notably theory papers, which we did not incorporate, for fear of making the list too long to be easily interpretable.

A key feature of our set up was that authors chose the paper type for their submission and therefore which review form would be used to

[4] http://coling2018.org/index.html%3Fp=156.html

[5] http://coling2018.org/index.html%3Fp=156.html

[6] For a quantitative view on this, we surveyed all authors shortly after the submission deadline, receiving 434 responses. We asked if it was clear to authors which paper type was appropriate for their paper and if they thought paper types are a good idea. 78.8% said it was clear and 91.0% said it was a good idea. (Interestingly, 74 people who said it wasn't clear which paper type was a good fit for theirs nonetheless said it was a good idea, and 21 people who thought it was clear which paper type fit nonetheless said it wasn't.)

evaluate it. Members of the program committee (including us) found plenty of papers that we thought were misclassified, but we decided against reclassifying any, because we couldn't conceive of a fair and consistent process for this and because we didn't want to second-guess authors. One frequent apparent misclassification was papers that appeared to us to be NLP engineering experiment papers submitted as computer-assisted linguistic analysis papers. This may have also been a result of cross-disciplinary confusion. Our description of the computer-assisted linguistic analysis type was *The focus of this paper type is new linguistic insight. It might take the form of an empirical study of some linguistic phenomenon, or of a theoretical result about a linguistically-relevant formal system.* It's possible that someone without training in linguistics would not know what terms like *linguistic phenomenon* or *formal system* denote for linguists, which in turn speaks to the need for more venues for interdisciplinary interaction.

A quantitative overview of COLING 2018 paper types is provided in Table 0.1. With the caveat that authors' conceptions of the different paper types didn't necessarily match ours, we were pleased to see that COLING 2018 was able to attract a broad range of papers and that reviewers felt similarly confident in reviewing all of them.

Paper	# Submissions	Avg Score	Avg Reviewer	Acceptance
Type			Confidence	Rate
NLPEE	657	2.86	3.51	37.94
CALA	163	2.85	3.42	33.33
Resource	106	2.76	3.50	32.32
Reproduction	35	2.92	3.54	48.57
Position	31	2.41	3.36	32.00
Survey	25	2.93	3.58	54.55
Overall	1017			37.27

Table 0.1: Quantitative overview of COLING 2018 paper types (NLPEE stands for NLP engineering experiment; CALA is computer-assisted linguistic analysis. Acceptance rate excludes papers that were withdrawn from the denominator.)

We ran a reviewer survey to get a sense of the process from a reviewer's perspective, particularly around paper types.[7] In general, the

[7] This was sent with some delay (on 25 May, though reviews were due 10 April) and, as some survey respondents pointed out, we may have gotten more accurate answers if we'd asked more quickly. The response rate was also relatively low: only 128 of our 1200+ reviewers answered the survey. No question was required, so the answers don't sum to 100%.

reviewers found that the authors had picked the correct paper type (69.5% responding “Yes, all of them” to the question “Did you feel like the authors chose the appropriate paper type for their papers?” and only one reviewer responding “No, none of them”) and that the review form questions were as good as or better than typical NLP conference review forms for evaluating papers of each type (when the papers were correctly assigned; 29.7% chose “Yes, better than usual for conferences/-better than expected”, 57% “Yes, about as usual/about as expected”, 6.3% “No, worse than usual/worse than expected”, and 1.6% “No, the review forms were poorly designed”).

We are pleased to see this concept live on in the call for papers for COLING 2020,[8] which invites NLP engineering experiment, computer-aided linguistic analysis, resource and reproduction papers. It has also been adopted by the Northern European Journal of Language Technology,[9] which invites the same paper types as COLING 2018.

3 A broad lens on quality of scholarship

The scholarship of our field is frequently criticized on three points: clarity of hypotheses, depth of analysis, and reproducibility. How can these be translated into clear signals in the submission and review process? In this section we discuss how we the dimensions of research quality that we wanted to emphasize into signals in the submission and review process. In each case, we provided guidance to authors via blog posts,[10] and offered in distilled form to authors and reviewers in the review forms. Finally, for reproducibility, we added further incentive in terms of the best paper process.

There are many ways that research can be presented an interdisciplinary field. This makes it hard to give generic advice on how to form and communicate research. Nevertheless, there are some essentials for making many types of research meaningful. We concentrated on three: hypotheses, analyses, and reproducibility.

[8] https://coling2020.org/pages/call_for_papers

[9] https://www.nejlt.org/authorinfo/; the authors are associated with this journal.

[10] We had a series of five guest blog posts on reproducibility, by Antske Fokkens, Liling Tan, Alice Motes, Kalina Bontcheva, Saif M. Mohammad. We published a post on error analysis and Fokkens’ post touched on this as well. The most valuable discussion around clarity of hypotheses came in the form of a comment by Bonnie Webber on our post about paper types.

We leveraged the instructions to authors, specifically in the form the paper types as described in the call for papers and the provision of the review forms for author inspection, to promote clarity of hypotheses. For engineering experiment (and some analysis) papers, a primary source of the value of the contribution is demonstration of a phenomenon. A clear hypothesis is critical to this and concentrates both the background and framing of the paper towards the essential question being asked. At COLING we directly requested authors give a clear hypothesis statement for empirical papers and gave a dedicated a review form point for assessing the quality of hypothesis statements.

We similarly used the call for papers and review forms to push for good analyses in papers. A good analysis tells us something about why method X is effective or ineffective for problem Y. It can tell us which are the more and less difficult parts of problem, directing future research and progress by concentrating effort more effectively onto unsolved areas. Analysis categories are not necessarily determined ahead of time, but rather emerge from the data. Does your sentiment analysis system get confused by counterfactuals? Does your event detection system miss negation not expressed by a simple form like *not*? At a superficial level, error analysis can involve looking at token frequencies or confusion matrices. A more advanced analysis could perform ablation experiments, to give signals about which parts of an approach have what impact over the evaluation data examples; examine system inputs to determine whether specific linguistic phenomena prove problematic for the algorithm; or use probing methodology to find correlations between a neural network structure and linguistic phenomena. Finally, a result is often more convincing if the hypothesis not only predicts an overall result, but also the kinds of errors or successes that an approach yields.[11]

Finally, we leveraged both the review forms and the best paper awards to incentivize reproducible research. Results that can't be consistently reproduced are not reliable; results that can be readily reproduced enhance understanding of a method and its use. More stringently, experiments that can be completely *replicated* using the original data and code have the potential to greatly advance understanding, by exposing precisely how results were derived, and sharing with others the materials needed to create them. Nevertheless, the proportion of empirical work in computational linguistics relaying sufficient methodological

[11] http://coling2018.org/slowly-growing-offspring-zigglebottom-anno-2017-guest-post/

detail to be reproduced is low [6], let alone the amount of work that provides for replication. At COLING 2018, we explicitly asked for reproduction and replicability assessment during review, where code and data had to be submitted for top scores; in contrast, promises of delivering these artifacts later were given the second-worst score. We openly excluded submissions that didn't include all relevant code from best paper award eligibility. As work with e.g. intellectual property (IP) restrictions might not include code, a notification was sent to authors of papers shortlisted for a best paper award, including this requirement. The intent was to give a direct incentive to release code before awards were given, rather than silently downgrading those with IP constraints. Nevertheless, at least one paper was de-selected for a best paper award for failing to meet this requirement. However, the general exercise was a success: in the end, a large number of papers were submitted with full code — around a third of all published papers across all types. This is a step towards reproducibility, though broad replicability of published results remains something the field struggles with.

4 Overcoming language bias

As COLING has had an international focus since its inception, it seemed particularly important to work to mitigate language bias in our field, on three levels: (1) bias against work on languages other than English; (2) bias towards English as the de facto language of scientific communication; and (3) bias against papers written by people with less fluency in English. We aimed to address (1) by specifically including work on different languages as a kind of novelty in our reviewing criteria for NLP engineering experiment and resource papers. Regarding (2), we continued the COLING practice, initiated by Martin Kay and Christian Boitet at COLING 2012, of inviting authors to include an abstract for their paper in a language of relevance other than English. This might be their own first language or a language under study in the paper. Finally, regarding (3), we instituted a writing mentoring program, which ran before the reviewing process, and is described in the remainder of this section.

The writing mentoring program was optional and was focused on helping those who perhaps aren't used to publishing in the field of computational linguistics, are early in their careers, and so on. We see mentoring as a tool that makes COLING accessible for broader range of high-quality ideas. In other words, it wasn't about pushing borderline papers into acceptance but rather alleviating presentational problems with papers that, in their underlying research quality, easily make the high required standard.

We advertised the program via the COLING PC blog[12] (as well as through the call for papers, distributed on various mailing lists and posted on the website[13]) and recruited prospective mentors from among the people signed up to be reviewers. Authors wishing to participate in the program were asked to submit an abstract four weeks ahead of the paper submission deadline. We assigned papers to mentors based on the abstracts, giving priority first to authors at non-Anglophone institutions and secondarily to any authors from institutions not yet well represented in international computational linguistics conferences. Authors then provided their drafts by three weeks ahead of the submission deadline and mentors provided feedback within one week, using a "mentoring form" created by the PCs and structured to encourage constructive feedback. This was done via the START system, but we encouraged mentors to provide contact information so that authors could get in touch with them if they had questions. We ensured that no mentor served as a reviewer for a paper they had mentored. Mentors were recognized in the COLING program,[14] but there was no indication of which papers received mentoring, either at the reviewing stage or at the publication stage.

We asked mentors to answer the following questions (refined in light of comments received on the PC blog) to structure their feedback:

- What is the main claim or result of this paper?
- What are the strengths of this paper?
- What questions do you have as a reader? What do you wish to know about the research that was carried out that is unclear as yet from the paper?
- What aspect of the paper do you think the COLING audience will find most interesting?
- Which paper category/review form do you think is most appropriate for this paper?

[12] http://coling2018.org/writing-mentoring-program/
[13] http://coling2018.org/index.html%3Fp=491.html
[14] We also recognized six outstanding mentors.

- Taking into consideration the specific questions in that review form, in what ways could the presentation of the research be strengthened?
- If you find places where there are grammatical or stylistic issues in writing, or in general, if you think certain improvements are possible in terms of overall organization and structure, please indicate these. It may be most convenient to do so by marking up a pdf with comments.

We also asked mentors to abide by a code of conduct, specifically agreeing to:

- Maintain confidentiality: Do not share the paper draft or discuss its contents with others (without express permission from the author). Do not appropriate the ideas in the paper.
- Commit to prompt feedback: Read the paper and provide feedback via the form by the deadline specified.
- Be constructive: Avoid sarcastic or harsh evaluative remarks; phrase feedback in terms of how to improve, rather than what is wrong or bad.

We were initially worried about having more demand for this program than we could support. However, in the event, we had over 100 potential mentors sign up and only about 50 requests for mentoring. In our author survey, we included questions whose aim was to find out if our authors were aware of the writing mentoring program and, for those who were but didn't take advantage of it, why not. 277 of 434 respondents (63.8%) said they were aware of it. The most common reason chosen for not taking advantage of it was "I didn't/couldn't have a draft ready in time." (150 respondents), followed by "I have good mentoring available to me in my local institution" (97 respondents). The other two options available in that check-all-that-apply question were "I have a lot of practice writing papers already" (74 respondents) and "Other" (10). Alas, a few people indicated that they only discovered it too late.

Leaving time for writing mentoring is definitely in tension with the just-in-time production of prose characteristic of our field. However, for a writing mentoring program to make the conference more accessible to authors with good ideas but less access to writing mentorship, it has to happen before the reviewing phase.[15] We are optimistic that if writing mentoring programs become a regular feature of conferences, people will incorporate them into their planning. This tension can be alleviated by encouraging nearly complete papers for the mentoring phase. That is, writing mentoring can be done effectively if the paper draft includes background, hypothesis, methodology, and the like, but still lacks final results.

5 Managing Anonymity

It is in the best interests of authors, conference attendees, and the field at large for reviewing to serve its gate-keeping function as fairly as possible. When reviewers have access to author identities while reviewing, biases come into play: Work from prestigious institutions gets more attention, with more reviewers bidding for it [12]. Knowing the gender of authors obscures the accurate assessment of research quality [4, 10, 13]. Prestigious institutions receive further favour, achieving better outcomes [1, 3, 9, 11, 12]. These factors are compounded when the volume of work accepted to an event is limited, either in terms of a fixed acceptance rate, or a fixed number of presentations. The latter factor's severity is shaped by the time and space available for the conference, and sometimes can only be raised at a cost to the participants. Limiting the amount of research presented, while offering a benefit to delegate decision making, also presents an opportunity cost: artificially limiting what can be presented excludes good work and, in the presence of biased reviewing, forces out excellent work by authors in some categories in favour of equally or less-excellent work by authors in other categories. This is inefficient for the field and unfavourable to conference attendees.

Traditionally, attempts to mitigate these biases focus on concealing the identity of some actors within the review process. Precisely which groups are able to see what varies depending on particular venues; the typical NLP/CL setup has been that author identity is concealed from reviewers and vice versa, but the rest is open (i.e. area chairs (ACs) can see who reviewers and authors are, and reviewers can identify ACs). This brings a variety of biases: Area chairs knowing author identity exposes authors to the same biases as found when author identities are

[15] From the PC point of view, using the conference management system to run the mentoring program a few weeks ahead of the submission deadline also served as an extremely valuable dry run with that software.

known to reviewers. In fact, the area chair has a large amount of influence on acceptance decisions, especially when acting without oversight. Reviewer identity being available to other reviewers and to ACs can also disrupt quality review: others can be disinclined to disagree with well-known names for fear of retribution, or be unconsciously (or consciously) swayed to weight reviews based on ethos rather than logos.

Our solution was to hide by default all the above identities. Author names were concealed, but so were reviewer and AC names, to the extent possible. Exceptions were: ACs to each other, reviewers to ACs/PCs, ACs/PCs to reviewers, and ACs and PCs to each other. All names were available to the general chair and PC co-chairs, but took some effort to retrieve; this enabled anonymous decisions to be made on borderline papers all the way to the top. Further, the names of authors of accepted papers were not released until the best paper committee had selected best papers — an important step if paper awards are to be free from author and affiliation biases.

6 Improving process consistency

Faced with the prospect of over 1000 submissions to the event, we built review processes to be consistent and fair across a large number of reviewers and area chairs (ACs). Here we describe three key features of our approach: having ACs work in teams, using dynamic areas (inspired by NAACL 2016), and communication.

A common theme in improving peer review is to establish practices that are fair. A large, complex event with submission counts in the thousands is liable to exhibit variation in reviewing standards — that is normal. But what can we do about the human factors involved in order to minimise this variation?

Area chairs make acceptance recommendations to program chairs, so this is probably the most important place in the process to have consistent decision making. Variation at this point can quickly lead to inconsistent decision making. At COLING, we worked with ACs to establish processes aimed at reduce variation here, while also improving the reviewing experience for all involved.

Individual AC bias can be moderated by having ACs work in pairs, jointly responsible for all decisions. This has multiple benefits: workload can be shared when life happens or during exceptional events;[16] pair work can have higher quality and efficiency [7]; and having two chairs gives a "wisdom of the crowds" effect, reducing the frequency and impact of extreme results from individuals.

Key to having high quality and consistent reviewing is balancing the load across area chairs. Rather than define areas by topic, and risk some being very large indeed, we followed the work of Ani Nenkova and Owen Rambow (PCs for NAACL 2016) and used dynamically defined areas. We asked ACs and reviewers to indicate their research topics (along dimensions such as task, methods, and languages). Based on this, we paired ACs according to experience, time zone compatibility, and research topics, and then matched reviewers to ACs based on research topics. The collective expertise of ACs and reviewers defined the dynamic areas. Submitted papers were then automatically allocated to areas according to the topics selected by authors at submission time, with areas being given the best-matching paper in round robin fashion, to give each area a roughly similar proportion of highly-relevant matches. The goal was to get people reviewing in their domain, rather than prescribe a specific structure of research topics. This departure from convention led to some confusion around area themes. For example as ACs and reviewers requested being allocated to "the machine translation area". With over 40% of papers, ACs and authors listing proficiency in machine translation, a single MT area would have been unwieldy.

On the other hand, the dynamic areas meant more even workload sharing and so more hope of a consistent reviewing process. Workload sharing worked well, or at least did not hamper performance: area reports came in on time and AC feedback was generally rapid.

Clear communication was the other tool used to move towards a consistent process. We provided an AC training manual, detailing all key events that involve ACs, and giving brief guides for each stage that covered what to expect, what is expected, and when the task is to be finished. Instructions needed to be concise and the guide sufficiently well structured for effective results from ad hoc access. Following air crew management principles [8], we encouraged ACs to intercede in the other's process and have communication around uncertainty. ACs

[16] We also had a small pool of "special circumstances" ACs, ready to step in at short notice should need arise.

were best-effort paired with other ACs in the same timezone to permit low-latency conversations.

We also created reviewing guidelines. Careers can depend on paper acceptances; at the least, papers represent an amount of work by the author that is best served with *formative* feedback. All reviews should meet this benchmark. Authors in recent years had commented that that review quality was a pain point; to engage with the community and take action on this, COLING review guidelines formed a code of conduct for reviewers with four main points: be timely, be constructive, be thorough, and maintain confidentiality.[17] ACs were asked to check reviews so that these guidelines were adhered to and to remind reviewers to improve their reviews in various cases, including very short reviews.

7 Acceptance criteria

Once the reviews are in, the next task is to turn reviewers' input into acceptance decisions. There was a good amount of space at the venue, so the concern became finding papers of high enough quality to publish, rather than allocating scant conference slots. The "loss function" for paper acceptance, i.e. the penalty for getting acceptance decisions wrong, is asymmetrical: include too many papers and the venue will not respect readers' time, making it less attractive — include too few and good work will be cut out [5], wasting time and also disinclining future submissions. In this section, we briefly describe our processes for determining acceptance, in light of the fact that the reviewing process is always noisy.

Two major constraints exist around paper acceptance: the volume of works that can be presented at one conference event and whether or not an individual paper is worth the audience's attention. The volume of works presented depends both on physical factors, around the size of the conference space, and on human factors, namely conference duration (one is worn down by long conferences) and the length of each conference day. COLING 2018's venue could accommodate many presentations, so focus was on how to determine acceptance.

We asked each pair of ACs to discuss papers in their area and produce recommendations based on the reviews and any author response. We invited authors to write brief responses to the reviews, addressed

[17] http://coling2018.org/index.html%3Fp=601.html

to (and shared with) the ACs only. This allows authors to flag mistakes in reviews while avoiding typically unproductive reviewer-author conflict. Authors often wonder if responding to reviews is worthwhile. We found that, for the majority of the papers which were accepted despite low reviewer score(s) (and correspondingly harsh reviews), our notes reflected effective author responses.[18]

We asked ACs to provide recommendations as a ranked list in four ordered segments: accept, maybe accept, maybe reject, reject. Due to reviewer subjectivity, both in terms of predisposition and accuracy, the worst way to put papers into these categories is by ordering by score. A sample of three reviewers is simply too weak. Confidence-weighted scores are no help; the meaning of confidence scores is not calibrated across reviewers, rendering it an unreliable multiplier. Rather, ACs need to consider scores in light of the review text and their own assessment of the paper. The amount of papers deemed maybe accept or maybe reject by the ACs varied greatly by area (min=0.0%, mean=13.8%, median=11.1%, max=41.4%, variance=13.2), with most areas placing a minority in these categories, and some just one or no papers at all, while others were unable to clearly decide on almost half.

In order to create final acceptance decisions, we divided the recommendations into “clear” and “borderline”. The borderline cases were either papers that the ACs marked as “maybe accept” or “maybe reject”, or, for areas that didn’t use those categories, the last two “accept” papers and the first two “reject” papers in the ACs’ ranking. This gave us a bit over 200 papers (19% of submissions) to consider. We divided the areas into two sets, one for each PC, making sure not to take papers with which we had COIs in our own stack, but otherwise keeping authors anonymous. Area by area, we looked at the borderline papers, considering reviews, reviewer discussion (if any), author response, AC comments, and sometimes the papers (to clarify particular points; we didn’t read the papers in full), and, in 23 cases of remaining uncertainty, discussed between ourselves to reach a decision.

In total, this process led to no individual having full power over a paper’s acceptance status at any step, thus reducing bias.

[18] We also provided reflections on what makes an author response effective on the PC blog.

8 Recognising excellent work

Best/outstanding paper awards have two purposes: On the one hand, they provide a chance for the conference program committee to highlight particularly compelling papers and promote them to a broader audience. On the other hand, they serve as recognition to the authors that may help advance their careers. However, when only a small number of such awards are presented, there is a tendency for them to be focused on papers of the most conventional types, leading to a further impression that good work in our field must fit a certain template. To mitigate this, we created a range of awards, to recognize excellence in each of our paper types, plus four additional categories which were not tied to specific paper types but rather foregrounded excellence in research practices we wished to promote: (1) best evaluation, for a paper that does evaluation very well, (2) most reproducible, where the paper's work is highly reproducible, (3) best challenge, for a paper that sets a new challenge, and (4) best error analysis, where the linguistic analysis of failures is exemplary. There was no "overall" best paper award, that is, no one of the outstanding papers was elevated above the others.

We also observed that reviewers in our field are extremely unlikely to nominate papers for best paper awards, making this process heavily biased to preferences of the few who do nominate. To increase the level of signal, we added a question to the review form asking reviewers if this paper was the strongest of the papers that they themself reviewed. This gave ACs more information from which to nominate best papers.

Forty-four papers were nominated by ACs for consideration by our best paper committee, which had 11 members. We created subcommittees of the best paper committee to consider each award, such that each award was considered by two committee members and most committee members worked on two award types. The exception is the "Best NLP engineering experiment" award, as that award type had the most nominations (being the most frequent paper type among our submissions). The committee members working on that type focused only on it. We instructed the best paper committee to be open to the possibility that some awards go unallocated (if warranted) and also that a paper end up with a different award than the one it was nominated for. In the event, the best challenge award went unallocated.

The best paper committee's process began just after author notifications were sent out. In order to preserve author anonymity in the best paper award selection process, we did not post the list of accepted papers until the best paper selection was done. Individual authors were free at this point to post their own information, but we trusted our best paper committee to not go hunting for it. Similarly, in order to preserve anonymity while maintaining the requirement that data/code be released for award eligibility (see §3), one of the PC chairs took on the work of verifying this for each paper after the committee made their recommendations but before they were awarded.

The nine papers receiving awards were presented in a plenary session at the conference. We also recognized the remaining nominated papers in the program. To underscore the principle that mode of presentation (poster v. oral) is not an indicator of paper quality, these "AC picks" were dispersed across both types of presentation. For those presented as posters, we drew attendees' attention by posting an image of chili peppers (appropriate to Santa Fe, NM) next to each such poster.

9 Conclusion

The project of curating papers for a large, interdisciplinary venue brings many interesting challenges and opportunities. In this paper, we have reported the approaches we took to several aspects of the process while serving as PC co-chairs for COLING 2018. We took the opportunity to foster interdisciplinary collaboration, to promote reproducibility, and to push-back against English-bias in our field. We addressed the challenges of managing a review process at the scale of 1,000 submissions (admittedly already small by 2022 standards but a challenge nonetheless), while maintaining review and decision process consistency. While some of the mechanics of our process were specific to the process of conference reviewing, we believe that a lot of the key ideas here are equally applicable to journal reviewing or the new ACL Rolling Review set up: specialized review forms for different paper types, protecting anonymity and working towards consistency across the review process, offering writing mentoring, encouraging abstracts and other written products in languages beyond English, and recognizing many different kinds of excellence in research.

In sharing these ideas and reflections, we hope to enable others to build on our work, but we are not writing this only for future program chairs and journal editors. We intend these discussions to be informative to junior scholars joining our field, both in terms of providing some insight into the (often perceived as opaque) process of peer review and in terms of shedding light on the high level of volunteerism and community spirit in our field — and prospects for joining with that to make a difference.

10 Acknowledgments

The authors thank Olga Zamareva and Angelina McMillan-Major for constructive comments on this manuscript, and all the community members who participated in forming and and giving feedback on COLING 2018.

Bibliography

[1] J. J. Bartko. The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):199–199, 1982.

[2] E. M. Bender, L. Derczynski, and P. Isabelle, editors. *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

[3] R. M. Blank. The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *The American Economic Review*, pages 1041–1067, 1991.

[4] A. E. Budden, T. Tregenza, L. W. Aarssen, J. Koricheva, R. Leimu, and C. J. Lortie. Double-blind review favours increased representation of female authors. *Trends in ecology & evolution*, 23(1):4–6, 2008.

[5] K. Church. Reviewing the reviewers. *Computational Linguistics*, 31(4):575–578, 2005.

[6] A. Fokkens, M. Van Erp, M. Postma, T. Pedersen, P. Vossen, and N. Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, 2013.

[7] J. E. Hannay, T. Dybå, E. Arisholm, and D. I. Sjøberg. The effectiveness of pair programming: A meta-analysis. *Information and Software Technology*, 51(7):1110–1122, 2009.

[8] R. L. Helmreich, A. C. Merritt, and J. A. Wilhelm. The evolution of crew resource management training in commercial aviation. *The international journal of aviation psychology*, 9(1):19–32, 1999.

[9] D. N. Laband and M. J. Piette. Does the" blindness" of peer review influence manuscript selection efficiency? *Southern Economic Journal*, pages 896–906, 1994.

[10] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, 2012.

[11] K. Okike, K. T. Hug, M. S. Kocher, and S. S. Leopold. Single-blind vs double-blind peer review in the setting of author prestige. *Jama*, 316(12):1315–1316, 2016.

[12] A. Tomkins, M. Zhang, and W. D. Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

[13] R. Van der Lee and N. Ellemers. Gender contributes to personal research funding success in the netherlands. *Proceedings of the National Academy of Sciences*, 112(40):12349–12353, 2015.