Double-blind peer-reviewing and inclusiveness in Russian NLP conferences

Andrey Kutuzov[1][0000−0003−2540−5912] and Irina Nikishina[2,3][0000−0003−4910−8568]

[1] University of Oslo, Oslo, Norway
andreku@ifi.uio.no

[2] Laboratory for Models and Methods of Computational Pragmatics, National Research University Higher School of Economics, Moscow, Russia

[3] Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia
Irina.Nikishina@skoltech.ru

Abstract. Double-blind peer reviewing has been proved to be pretty effective and fair way of academic work selection. However, to the best of our knowledge, nobody has yet analysed the effects caused by its introduction at the Russian NLP conferences. We investigate how the double-blind peer reviewing influences gender and location (according to authors' affiliations) biases and whether it makes two of the conferences under analysis more inclusive. The results show that gender distribution has become more equal for the Dialogue conference, but did not change for the AIST conference. The authors' location distribution (roughly divided into 'central' and 'not central') has become more equal for AIST, but, interestingly, less equal for Dialogue.

1 Setting the question

Double-blind peer-reviewing means that the authors of the submitted papers do not know the names of the reviewers, and the reviewers do not know the names of the authors.

Peer review originates from the publishing process of Philosophical Transactions journal in the middle of the eighteenth century: its reviewing policy implied sending manuscripts to experts before publishing [3]. By the middle of twentieth century, peer reviewing has become the widely acknowledged standard for all top-tier international journals and conferences. Despite the long history, first papers devoted to single-blind and double blind review comparison date back only to the 1980s [4].

In comparison to the double-blind system, other setups where reviewers know the names of the authors have an obvious shortcoming: human subjectivity. In other words, the reviewers' decisions are inevitably biased (consciously or unconsciously). For instance, according to [5], papers by well-known authors are accepted 1.5 times more often, by well-known companies — 2 times more often, with a female first author — 20% less often. Double-blind peer reviewing successfully tackles the problem, significantly alleviating the bias [1]. Apparently, it does not solve all the existing issues, but it does make the scientific program more diverse and the conference itself more inclusive.

The main conferences in Computational linguistics and Natural Language Processing in Russia also try to keep up with that trend: AIST[1] switched completely to double-blind reviewing starting from 2017, Dialogue[2] did the same in 2019.

In this paper we set to find out whether the 'double-blind turn' influences the most widespread biases about gender and place of origin of the authors of the accepted papers. In particular, our research questions are:

1. Did the ratio of female authors in the accepted papers increased after the introduction of double-blind reviewing?
2. Did the number of 'non-centrally located' authors increased after the introduction of double-blind reviewing?

2 Inclusiveness in Russian NLP conferences

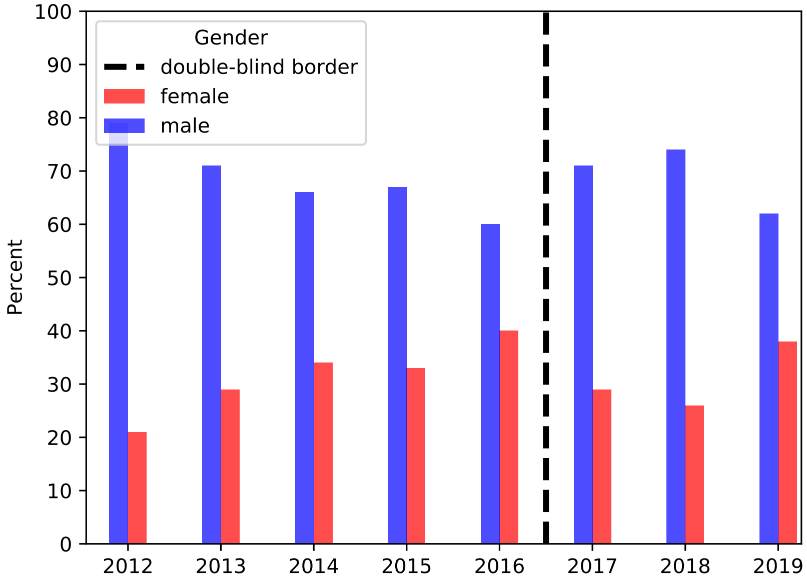

Fig. 1. Gender distribution among AIST authors from 2012 to 2019

[1] https://aistconf.org/

[2] http://www.dialog-21.ru/en/

In order to measure inclusiveness, we use data from the RusNLP project[3] [2] for the annual AIST (years 2012-2019) and Dialogue (years 2000-2019) conferences. Gender and geography metadata was annotated manually, based on authors' names and affiliations. Unfortunately, the RusNLP database does not contain information about the order of the authors, so we counted all of them equally. While annotating the 'location' or 'city' attribute, we used the following notation: 'centre' stands for Moscow, Saint-Petersburg and authors from outside Russia, while 'province' stands for all other regions and cities in Russia. We then calculated the percentage of male-female and central-provincial authors for each year and each venue.

In order to measure the difference in these percentages before and after the introduction of double-blind reviewing for AIST, we applied the Welch T-test. For Dialogue, we have only one data point with the double-blind reviewing (year 2019), thus the T-test is ill-defined, and we simply checked whether the absolute difference between the percentages exceeds the standard deviation of the respective values for the years before the double-blind introduction (2000-2018).

2.1 Gender distribution

Starting with AIST, we first calculated the average percentage of its female authors before and after double-blind introduction (31 before and 31 after). Obviously, there is no statistically significant difference here, according to the Welch T-test: $statistic = -0.08$, $P = 0.94$. The yearly percentages are visualised in Figure 1. In this and all the following plots, the dashed vertical line denotes the year after which the venue switched to the double-blind process.

The picture is different for the gender distribution of the Dialogue authors. Before the double-blind peer review, on average 57% of authors were males and 43% were females. This changed to 45% and 55% respectively in 2019 (see Figure 2. Thus, the ratio of female authors increased by 12 points. Naively comparing this value to the standard deviation of the yearly female percentages before the introduction of double blind reviewing (it is 5 across 18 years), we observe that the increase value exceeds the standard deviation more than two times. From this, we conclude that the difference is significant, and the number of female authors has indeed increased.

2.2 City/location distribution

For AIST, the average yearly percentage of 'central' authors before double-blind reviewing was 79%, but after introducing it in 2017, this value fell to 56%. The Welch T-test confirms that the difference is statistically significant: $statistic = 2.48$, $P = 0.048$. 'Provincial' authors indeed benefited from the double-blind process. This can also be clearly seen in Figure 3, which additionally shows the geographical location of the conference itself in each respective year ('E-burg' stands for Ekaterinburg). Interestingly, the introduction of double-blind

[3] https://nlp.rusvectores.org

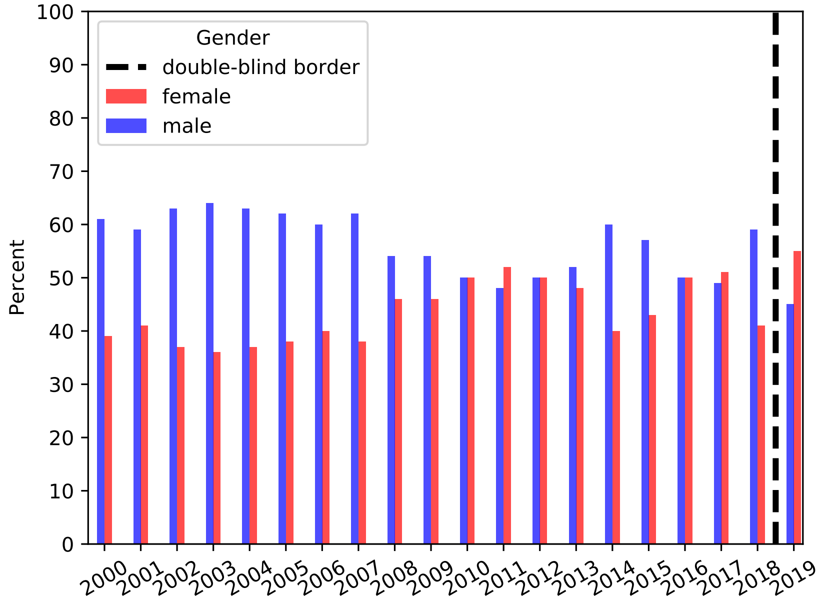

Fig. 2. Gender distribution among Dialogue authors from 2000 to 2019

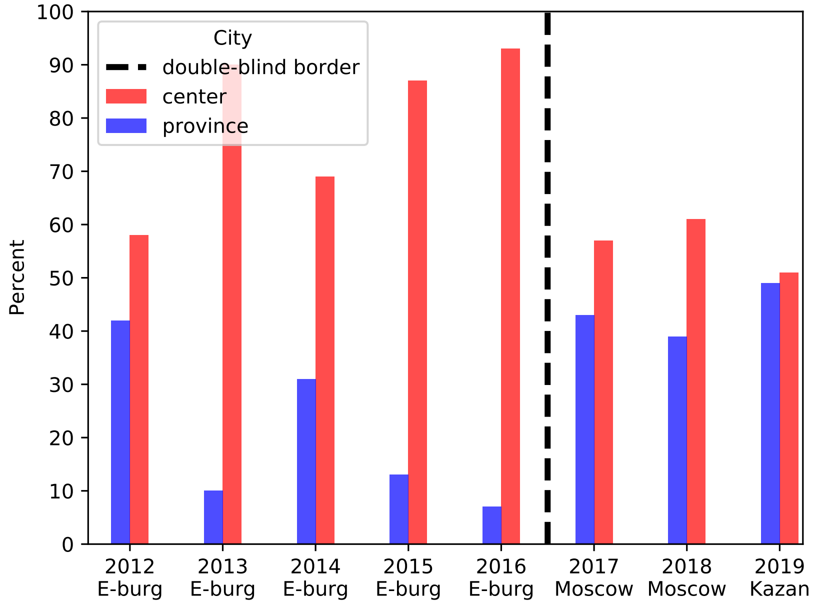

Fig. 3. City distribution among AIST authors from 2012 to 2019

reviewing has significantly decreased the ratio of ‘central’ authors, even though at the same time the conference itself moved to Moscow (years 2017 and 2018). This additionally confirms the significance of the discovered trend.

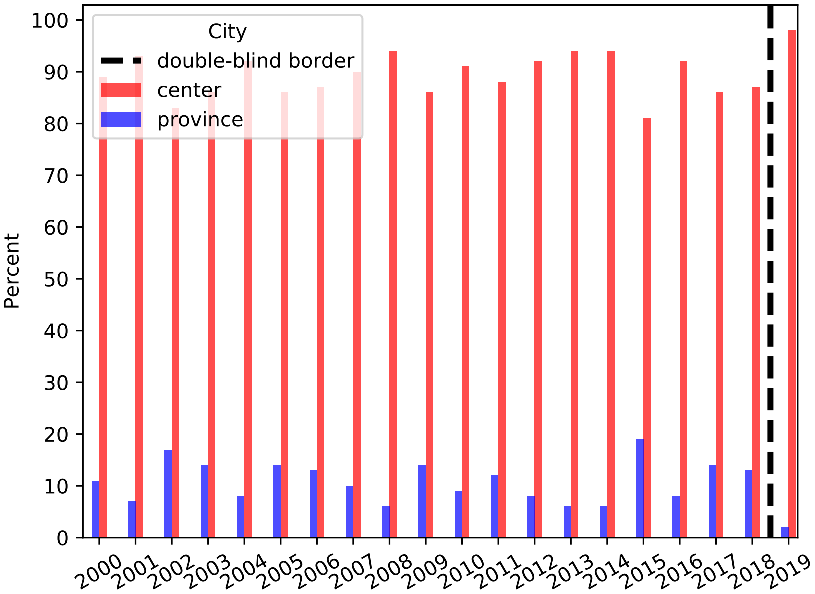

Fig. 4. City distribution among Dialogue authors from 2000 to 2019

For Dialogue, there were on average 89% of ‘central’ authors before and 98% after the introduction of double-blind reviewing. Thus, the percentage of ‘provincial’ authors has actually *decreased* from 11% to 2% (see Figure 4). This absolute difference of 9 is much higher than the standard deviation from the years 2000-2018 (4), so it seems to be significant. We believe this is caused by a random fluctuation: in Dialogue-2019, there is actually only *one* accepted paper authored by persons not located in Moscow, Saint-Petersburg or outside of Russia. This single paper is responsible for all the 2% of ‘provincial’ authors that we observe. It is of course difficult to do any conclusions on such an anecdotal evidence. More observations are certainly needed in the years to come.

3 Results

There are several significant changes occurring after introducing double-blind reviewing for both AIST and Dialogue conferences. First of all, there are no changes found in gender distribution of authors for AIST, which we guess means it has been egalitarian enough in this respect from the very start. At the same time, for Dialogue, we observe a significant increase in the ratio of female authors.

In case of location distribution, the results are rather controversial: there is a significant increase in the ratio of 'provincial' authors for AIST, but for Dialogue, there is a significant *decrease*. However, since the Dialogue statistics after the introduction of double-blind reviewing is based on a single observation, it is not fully reliable. We certainly have to wait until the year 2020 to see the forthcoming trends.

Another limitation of this pilot research is that obviously other factors may influence the distribution changes, e.g. conference location, or topical popularity. We have yet to find out how to exclude the influence of such extra factors. Finally, in the future we plan to expand our analysis by including the data from the AINL conference[4].

And of course, we welcome everyone to submit to all of the conferences mentioned above.

Acknowledgements The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'

References

1. Budden, A.E., Tregenza, T., Aarssen, L.W., Koricheva, J., Leimu, R., Lortie, C.J.: Double-blind review favours increased representation of female authors. Trends in ecology & evolution **23**(1), 4–6 (2008)
2. Nikishina, I., Bakarov, A., Kutuzov, A.: RusNLP: Semantic search engine for russian NLP conference papers. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 111–120. Springer (2018)
3. Spier, R.: The history of the peer-review process. Trends in Biotechnology **20**(8), 357–358 (2002)
4. Surwillo, W.W.: Anonymous reviewing and the peer-review process. American Psychologist **41(2)**, 218 (1986)
5. Tomkins, A., Zhang, M., Heavlin, W.D.: Single versus double blind reviewing at WSDM 2017. arXiv preprint arXiv:1702.00502 (2017)

[4] https://ainlconf.ru/