NLP Community Perspectives on Replicability

Margot Mieskes
University of Applied Sciences, Darmstadt
Germany
margot.mieskes@h-da.de

Karën Fort
Sorbonne Université, EA STIH
Paris, France
karen.fort@sorbonne-universite.fr

Aurélie Névéol
LIMSI, CNRS,
Université Paris-Saclay
France
neveol@limsi.fr

Cyril Grouin
LIMSI, CNRS,
Université Paris-Saclay
France
cyril.grouin@limsi.fr

Kevin Cohen
Computational Bioscience Program
University of Colorado, USA
kevin.cohen@gmail.com

Abstract

With recent efforts in drawing attention to the task of replicating and/or reproducing[1] results, for example in the context of COLING 2018 and various LREC workshops, the question arises how the NLP community views the topic of replicability in general. Using a survey, in which we involve members of the NLP community, we investigate how our community perceives this topic, its relevance and options for improvement. Based on over two hundred participants, the survey results confirm earlier observations, that successful reproducibility requires more than having access to code and data. Additionally, the results show that the topic has to be tackled from the authors', reviewers' and community's side.

1 Introduction

"As a community, we need to know where our approaches fail, as much – if not more so – as where they succeed." Despite this statement by Fokkens et al. (2013), we are still aiming at higher, faster, better results with little outside verification. And although it has become good practise to share code, data and parameters, previous work and experience indicate that sharing is still not as common as one would hope for. Call for Papers in major conferences encourage to submit or reference data and to submit code, i.e., in supplemental material[2]. But recent work indicates that this is not done thoroughly enough (Mieskes, 2017; Cohen et al., 2017; Wieling et al., 2018). Other factors mentioned by Fokkens et al. (2013), such as preprocessing methods, experimental setup, system variation, etc., are rarely reported. Pedersen (2008) urges to not fear any dispossession of a tool, and highlights that sharing code will result in its use in new systems and citations of the work describing this code. Moreover, we should consider sharing software as a way to improve it more efficiently. In recent years, we have seen a rise in attention targeted towards the task of replication for example in the COLING 2018 selection criteria[3], the LREC 4REAL Workshops (Branco et al., 2016, 2018) and the recent LREC initiative for replication.[4] But so far the task of replicating previous results has little merit in itself, but is rather only a (baseline) part in a paper. Additionally, normally it only gets reported in successful or mainly successful cases. Following the argument by Fokkens et al. (2013), the cases where it fails, hardly ever get reported, despite Calls for Papers encouraging negative results[5] and although they might be equally or even more important than the successful replication.

Our contribution therefore is to identify how the community views the topic of replication and what role each individual plays as an author, as a reviewer and as part of the NLP community. In conducting a survey, which drew answers from over two hundred respondents, we get a better picture of the factors that support or hinder making repli-

[1] We use *replication* to describe related efforts, regardless of the exact aim (see (Cohen et al., 2018))

[2] see for example http://www.acl2019.org/EN/call-for-papers.xhtml "ACL (...) encourages the submission of supplementary material to report (...) details necessary for the replication of the experiments."

[3] See: https://coling2018.org/paper-types/.

[4] This call occurred 5 weeks after we posted our call and are unrelated, but the latter might be inspired by our survey, see http://wordpress.let.vupr.nl/lrec-reproduction/.

[5] "A negative result" http://www.acl2019.org/EN/call-for-papers.xhtm (Short Papers)

Proceedings of Recent Advances in Natural Language Processing, pages 768–775,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_089

cation more visible and how the three factors described above influence this. Our results indicate that the participants in general regard replication as an important issue and that the NLP community could do more to support replication, which would strengthen the field as a whole.[6]

2 Related Work

One of the earliest reports of a replication effort addresses manual word-sense disambiguation based on four words, representing different degrees of difficulty (Kilgarriff, 1999). The author reports that humans agree in this task on average in 95% of the cases. Following Fokkens et al. (2013), others look into parameters that influence the replicability of results. Dakota and Kübler (2017); Marrese-Taylor and Matsuo (2017) and Horsmann and Zesch (2017) report various parameters and problems with replication experiments for morphology and syntax.

In the field of biomedical NLP, Olorisade et al. (2017) assess the reproducibility of findings published in 33 papers. They notice that data sets were missing, making it impossible to reproduce results for 80% of the papers. These figures are in line with results reported by Mieskes (2017). They consider that a permanent link to the resources (data set, software, etc.) must exist along with published papers. As part of a NLP challenge, Névéol et al. (2016) report results on replicating experiments from three systems submitted to the CLEF eHealth track. They show that replication is feasible although "ease of replicating results varied". They suggest the allowance of extra pages for papers, where information required to replicate an experiment could be reported.

Moore and Rayson (2018) illustrate how to publish relevant details to reduce efforts in repeatability and generalisability. Suggestions include using only open data, open source code and providing extensive documentation in the code.

Wieling et al. (2018) describe one example where exact replication was possible and the authors list the parameters that allowed them to do so: a virtual image, containing all code and all data or providing CodaLab worksheets. Their study, which compared the situation in sharing research artefacts between 2011 and 2016 indicates that, while the situation has improved and the availability of data is high, access to code is less so and requesting code is unsuccessful in most of the cases. Based on results of their actual replication experiments, *at most 60%* of the studies are replicable, but only if the need for exact replication was relaxed.

Fares et al. (2017) present a repository and infrastructure containing texts, tools and embeddings for English and Norwegian. Their aim is to facilitate replicability and testing of previous results. Dror et al. (2017) propose a "replicability analysis framework" and demonstrate its use on various tasks such as part-of-speech tagging or cross-domain sentiment classification. They specifically target cases where algorithms are compared across multiple data sets. The results indicate that testing on a range of data sets is only beneficial if the data sets are heterogeneous.

3 Survey Design

Our survey has 18 questions, of which many were conditional and show only if they apply to the respondent. Thus, not all questions have been answered by all participants, while most multiple-choice questions allow for several answers, resulting in more answers than participants for these questions. Questions are grouped into three categories: (i) replication work in general, (ii) replicating one's own work and (iii) replicating others' work.

General questions quiz participants on their perception of replication work. We also inquire about their current position to investigate potential correlation with other aspects of the survey. Questions addressing participants' replication experience specifically enquired about research artefacts availability (data, code, parameters, etc.) and about the timeline of the replication experience in order to assess attrition.

The survey was advertised on professional mailing lists (BioNLP, Corpora, LN and GLCL[7]) and social network (LinkedIn). The appendix gives details on the progression of responses. With respect to sensitive data, only e-mail addresses were provided, on a voluntary basis, and we follow the ACM Code of Ethics and Professional Conduct[8],

[6]The complete results of the survey are available at https://github.com/replicateNLP/Survey-RANLP2019. Please note, that due to privacy regulations, we had to remove some free text answers that contain personal information, such as E-Mail addresses, which were given on a voluntary basis.

[7]Biomedical NLP, French and German NLP.

[8]https://www.acm.org/code-of-ethics

specifically sections 1.6 and 1.7.

Figure 1 illustrates the flow of the questionnaire. If a person did neither replicate their own or someone elses work, they only had to answer the blue marked questions. If a participant tried to replicate his/her own work, but not someone elses work, they only had to answer the blue and the orange questions. Only persons who have experience in replicating their own and someone elses work had to go through the whole questionnaire, including the green questions.[9] This flow in addition to the possibility to give more than one answer in some questions results in different numbers of answers for each question.

4 Results

We received 225 responses and the two biggest groups of participants in our study identify themselves as graduate students and postdocs.

With respect to when work on replication has been done, 36 participants (16%) gave more than one anwer, indicating that they did work on replication at various stages of their career. Most answers (50.3%) state that replication was done on MSc or PhD level, less on PostDoc level (20.7%) and slightly more as Faculty members (24.3%). However, we did not find strong correlations between the respondents' position and opinion on the importance (or lack thereof) of reproducibility. Figure 2 shows the absolute numbers for this question, while Figure 3 shows the numbers for the participants current position.

4.1 General Stance towards Replicability

The answers show that in 56.4% of the cases, work on replication is considered "Important" and another 7.5% state that it is "Somewhat Important". Only 2 answers indicate that this work is "Unimportant". 20% of the answers regard work on replication as publishable, while 11.8% deem it unpublishable. 87 participants gave more than one answer. The majority (49) consider work on replication as important *and* publishable, while 26 of them consider it important *but not* publishable (see Figure 4 for the absolute numbers).

4.2 Replicating one's Own Work

Roughly 70% (156) of the participants declare they have tried to replicate their own work while about 30% have not tried (total 225).

[9]Please note that only the most important questions are illustrated here and some questions have been left out.

With respect to how often replication of the same experiment was tried, 15 participants gave more than one answer, giving us 172 answers (see Figure 5 for the detailed figures). Of these, 28.5% indicate that replication was tried only once, while 38.4% tried 3 times or more.

When looking at the last attempt (total answers 234), nearly half (47.0%) report that they reached the same general conclusions, while 23.1% state that they reached the same figures. A little less than 10% report that they managed to re-implement the system, but got significantly different results. Another 14.9% could not find either the code or the data or the parameters used for the experiment (see Figure fig:resultsOwn for the absolute numbers). 52 participants gave more than one answer of which 25 report that they reached the same general conclusions *and* the same figures. Overall, the results indicate that even in the case when researchers try to replicate their own work, they fully succeed in only 23.1% of the cases.

4.3 Replicating Others' Work

About 60% (total 130) of the participants report that they tried to replicate someone else's work (see Figure 7 for detailed, absolute numbers). 51 respondents gave more than one answer with respect to the results achieved when replicating somebody else's work, resulting in 211 answers. Approximately 40% of the answers state that they reached the same general conclusions or figures, while another 33.6% of the answers state that they managed to re-implement or re-run the system, but with significantly different results. Nearly 23.1% of the responses state that re-implementation or re-runnning of experiments was not achieved (see Figure 8 for the absolute responses for this question). This means that over half of the replication experiments failed, either early on or at the level of results achieved.

4.4 Accessibililty of Research Artefacts

For finding research artefacts such as code, data and parameters, respondents gave several answers, resulting in 250 answers for where the code can be found, 260 answers for finding data and resources and 233 answers for finding the experimental parameters. GitHub is by far the most popular (36.4% of the answers) for accessing *code*, but more than 23.6% of the answers state that code is found on the authors' personal webpage,

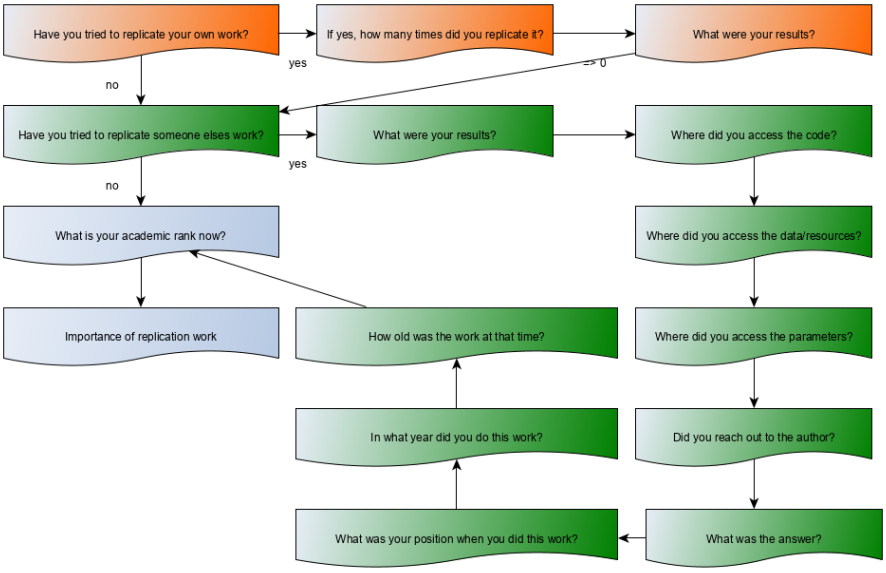

Figure 1: Illustration of the questionnaire flow.

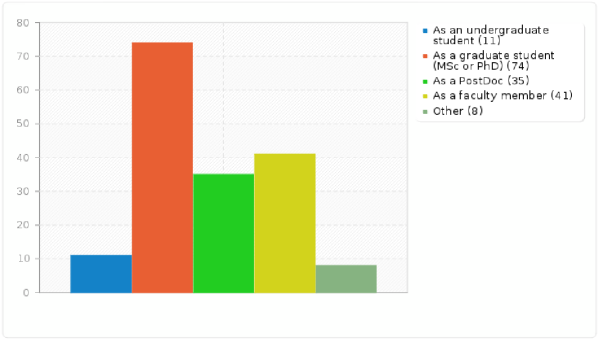

Figure 2: Position of the participants at the time they were doing replication experiments.

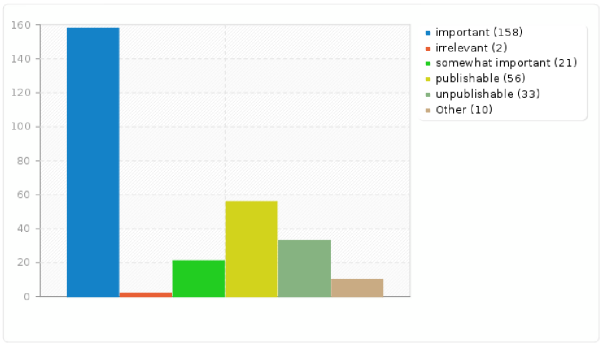

Figure 4: Importance of Replication in General.

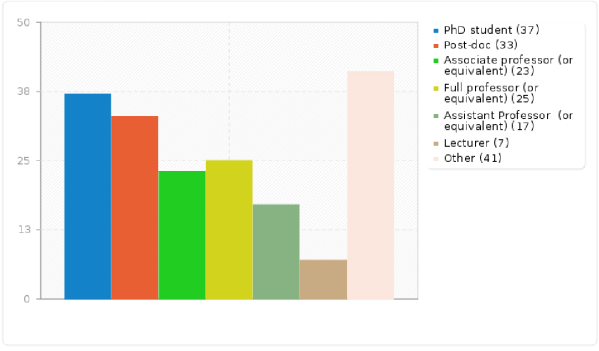

Figure 3: Position of the participants at the time of the survey.

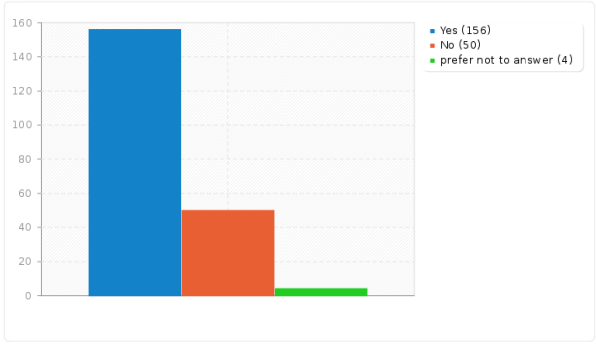

Figure 5: Participants who tried to replicate their own work.

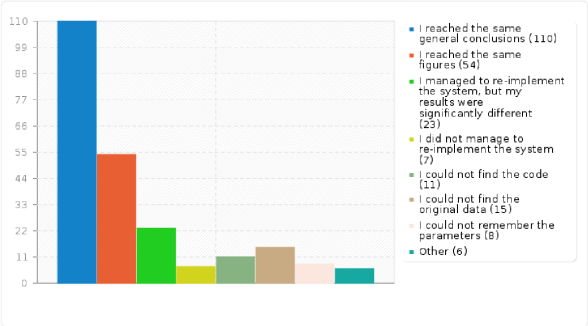

Figure 6: Results achieved when trying to replicate own work.

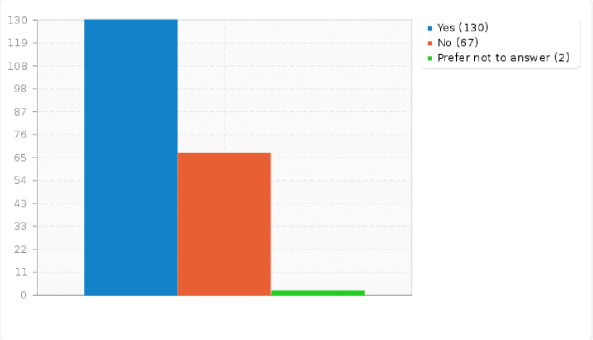

Figure 7: Participants who tried to replicate someone elses work.

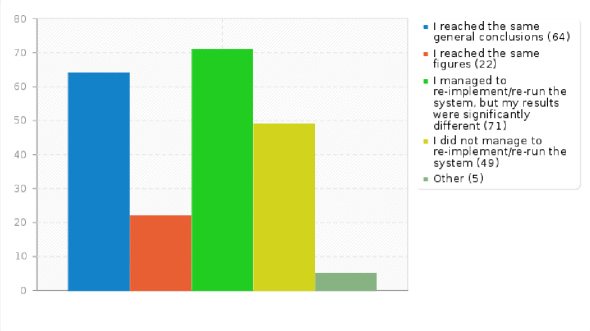

Figure 8: Results achieved when replicating someone elses work.

which does not guarantee availability beyond that person maintaining his/her webpage. More than 14% of the answers report that the code could not be found. *Data* is also primarily published via GitHub or personal webpages (25% and 25.7% of the answers respectively). 11.1% report that the material used for the experiments could not be found. *Parameters* for experiments are primarily found in the respective publications (40.3% of the answers), while 21.9% of the answers state they could be found on GitHub as well. 13.7% report that they could not find parameters at all. Figure 9 illustrates the absolute numbers concerning the availability of code, while Figure 10 and Figure 11 illustrate them for other resources and parameters respectively.

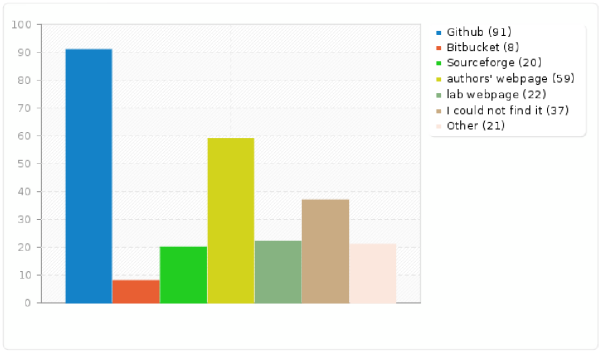

Figure 9: Sources for Accessing the Code for replication.

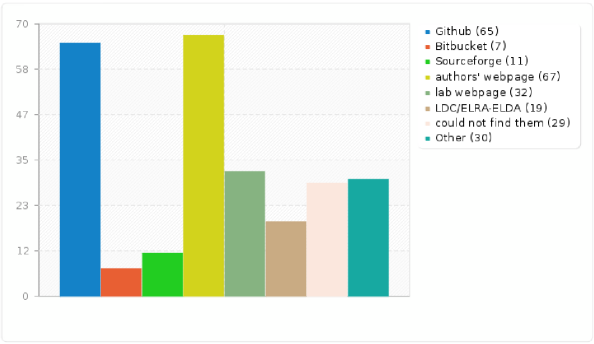

Figure 10: Sources for Accessing Data and other Resources.

Concerning these three elements, the text box associated with "Other" very frequently mentions "personal communication" or "e-mail" as a way of obtaining necessary information. This gives rise to a range of further issues, legal, ethical and in terms of transparency. Besides, it is only possible if authors actually answer such e-mails.

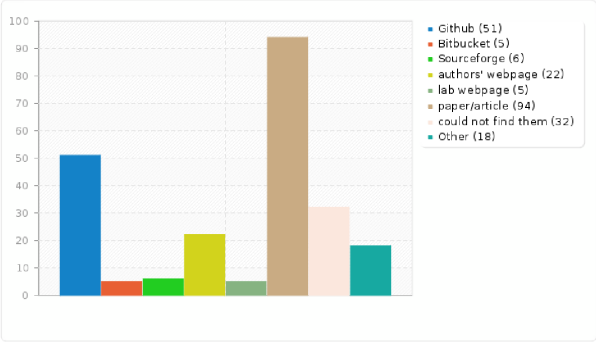

Figure 11: Sources for Accessing Parameters for Replication.

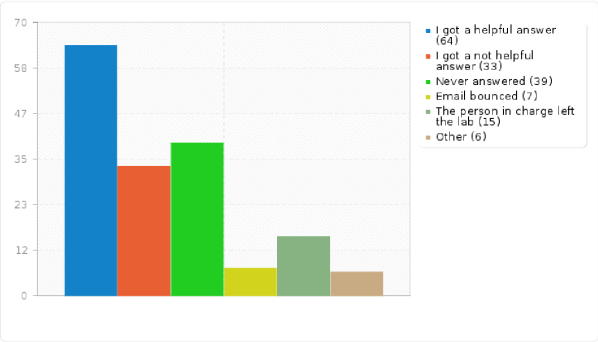

Figure 13: Quality of Answers by original authors.

In our survey, 40% of the participants report that they tried to get in touch with the authors (see Figure 12), but only in about 30% of the reported cases received a helpful answer (164 answers received, as 41 participants gave more than one answer). Approximately 20.1% mention that unhelpful answers were received and almost 23.8% never received any answer. Due to evolution of careers, 13.4% found out that the person had left the lab or the e-mail bounced, resulting in almost 40% of examples where authors were unreachable (see Figure 13).

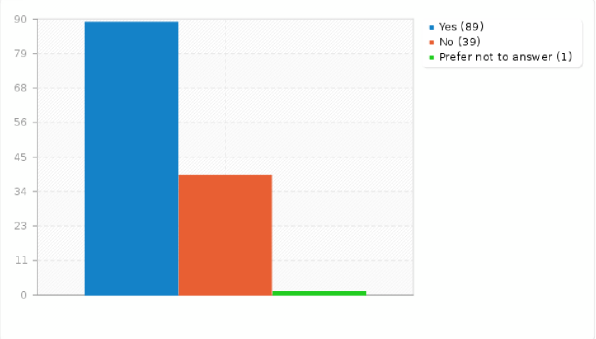

Figure 12: Participants who state that they reached out to original authors.

5 Discussion

The survey results suggest that **replicability is perceived as an important issue** by a majority of responses ($\geq$ 60%). It is difficult to compute an accurate response rate for the survey, because we do not know the extent of the population comprised by subscribers to the mailing lists and professional networks that we reached out to. However, if we approximate the target population using the average participation in a *ACL conference (N=1,000), we can estimate the response rate at about 20%. This has previously been described as an "acceptable" response rate for online surveys.[10] While the responding population includes researchers with a wide range of seniority as well as members of academia and industry, it could be biased by their interest in replication. The results, and especially the comments, indicate that this is most likely not the case as some participants do not see the value in replicating previous results. They state that replicating previous work is only an "exercise" and actually an "overload" on the already busy researchers' schedule. One comment states that *"10 year old systems are irrelevant"* and *"ML is moving so fast"*, which renders replicability studies essentially worthless. This could explain why the two largest groups performing replicability studies are on PhD or PostDoc level.

What does this mean for replicability? Responses indicate a variety of views on how replicability should be facilitated. One comment states that there is already a culture of sharing and publishing data and code and this should be enough. Another participant states that in industrial research publishing data or code is difficult, but suggests that the validity of an approach can be proven by applying the method to other data and in reaching the same general conclusions. Some participants support the idea of giving more visibility to replicability by making it a prominent topic at major conferences and by enforcing reporting guidelines towards reproducibility.

Some participants mention that replicability is crucial to be taken seriously as a scientific field, even stating that the field is "suspect" if replication fails. One participant suggests that *"every paper*

[10] http://socialnorms.org/what-is-an-acceptable-survey-response-rate/

cited enough times should be replicated".

6 Conclusions and Future Work

Based on a survey we gave insight into the NLP community's view on replicability. We targeted three different facets of this topic: Authors publishing their work, Researchers building on top of other researchers' work and the Community, supporting such efforts. Our results show that on the authors' side more information has to be shared openly, rather than via personal communication. The use of reporting guidelines formalized into a protocol has been suggested recently for clinical NLP (Velupillai et al., 2018). Earlier studies from the clinical domain suggest that adherence to such guidelines is suboptimal (Samaan et al., 2013) and methods to improve adherence are being investigated (Blanco et al., 2017). The task of creating guidelines falls to the community and the adherence of such guidelines could become part of the reviewing process.

Experiments in replication fail more often than not. If we document and store all relevant information so that results could be reproduced by ourselves (e.g., before the final paper submission), the package could be published completely. Results by Wieling et al. (2018) indicate that images containing all the material or technical lab books published on CodaLab might be a way to proceed. Additionally, failure to replicate previous work (i.e. not achieving the same results and/or not being able to draw the same conclusions as previously reported), should be publishable in a way that gives us scientific merit and could be encouraged more.

Based on the comments, each of us, in all of our individual roles can improve the situation: As authors, we can be more diligent when reporting our experiments and experimental setup—even testing the replicability of our experiments ourselves. As reviewers, we can be more careful to check the supplementary material for relevant information, pointing out missing elements. As a community, we can appreciate replication more and develop guidelines both for authors and for reviewers.

Future Work The next steps include, but are not limited to, analyzing whether the supplementary material and appendices actually do improve replicability, as stated by Névéol et al. (2016). Furthermore, evaluating the repository offered by Fares et al. (2017), whether other researchers actually build on top of it and with what results. NAACL recently initiated a "test of time" award. This could be extended to consider experimental work that has been cited often and gained influence on a shorter time-scale. This work could be verified for a follow-up conference. Replicability could also become a factor in the best paper awards.

References

David Blanco, Jamie J Kirkham, Douglas G Altman, David Moher, Isabelle Boutron, and Erik Cobo. 2017. Interventions to improve adherence to reporting guidelines in health research: a scoping review protocol. *BMJ Open*, 7(11).

António Branco, Nicoletta Calzolari, and Khalid Choukri, editors. 2016. *Proceedings of the Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*.

António Branco, Nicoletta Calzolary, and Khalid Choukri, editors. 2018. *Proceedings of the 4REAL 2018 - Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*. European Language Resources Association, Paris.

K. Bretonnel Cohen, Aurélie Névéol, Jingbo Xia, Negacy Hailu, Larry Hunter, and Pierre Zweigenbaum. 2017. Reproducibility in Biomedical Natural Language Processing. In *AMIA annual symposium proceedings*, page 1994. American Medical Informatics Association.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three Dimensions of Reproducibility in Natural Language Processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA).

Daniel Dakota and Sandra Kübler. 2017. Towards Replicability in Parsing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017* Warna, Bulgaria, 2–8 September 2017, pages 185–194. INCOMA Ltd.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic*

Conference on Computational Linguistics Gothenburg, Sweden, 22–24 May, 2017, pages 271–276. Association for Computational Linguistics.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In *Proceedings of the 51st Conference of the Association for Computational Linguistics* Sofia, Bulgaria 4–9 August 2013, pages 1691–1701.

Tobias Horsmann and Torsten Zesch. 2017. Do LSTMs really work so well for PoS tagging? – A replication study. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* Copenhagen, Denmark, September 9–11, 2017, pages 727–736. Association for Computational Linguistics.

Adam Kilgarriff. 1999. 95% Replicability for Manual Word Sense Tagging. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-1999)* Bergen, Norway, 8–12 June, 1999. Association for Computational Linguistics.

Edison Marrese-Taylor and Yutaka Matsuo. 2017. Replication issues in syntax-based aspect extraction for opinion mining. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics* Valencia, Spain, April 3–7, 2017, pages 23–32. Association for Computational Linguistics.

Margot Mieskes. 2017. A Quantitative Study of Data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–7, Valencia, Spain. Association for Computational Linguistics.

Andrew Moore and Paul Rayson. 2018. Bringing replication and reproduction together with generalisability in NLP: Three reproduction studies for Target Dependent Sentiment Analysis. In *Proceedings of the 27th International Conference on Computational Linguistics* Santa Fe, New Mexico, USA, August 20–26, 2018, pages 1132–1144. Association for Computational Linguistics.

Aurélie Névéol, K. Bretonnel Cohen, Cyril Grouin, and Aude Robert. 2016. Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis* Austin, Texas, November 5, 2016, pages 78–84. Association for Computational Linguistics.

Babatunde Kazeem Olorisade, Pearl Brereton, and Peter Andras. 2017. Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist. *Journal of Biomedical Informatics*, 73:1 – 13.

Ted Pedersen. 2008. Empiricism Is Not a Matter of Faith. *Computational Linguistics*, 34(3):465–470.

Zainab Samaan, Lawrence Mbuagbaw, Daisy Kosa, Victoria Borg Debono, Rejane Dillenburg, Shiyuan Zhang, Vincent Fruci, Brittany Dennis, Monica Bawor, and Lehana Thabane. 2013. A systematic scoping review of adherence to reporting guidelines in health care literature. *Journal of Multidisciplinary Healthcare*, 6:169–88.

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D. Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, and Rina Dutta. 2018. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88:11 – 19.

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.