Lang Resources & Evaluation (2017) 51:1–5
DOI 10.1007/s10579-017-9380-0

EDITORIAL

Replicability and reproducibility of research results for human language technology: introducing an *LRE* special section

António Branco[1] · Kevin Bretonnel Cohen[2] · Piek Vossen[3] · Nancy Ide[4] · Nicoletta Calzolari[5]

Published online: 16 February 2017

Investment in new research and technology is made under the assumption that scientific claims are supported by solid evidence; however, recent studies have shown that this is often not the case. For example, it has been shown that for some published results with major impact, replication of published results is difficult or impossible (e.g. Prinz et al. 2011; Begley and Ellis 2012; Fokkens et al. 2013; Anderson et al. 2015), and that exaggerated and false claims, sometimes with fabricated data and fake authors, have been accepted by and published in respectable journals (e.g., Fanelli 2009; Ioannidis 2011; Bohannon 2013; Hvistendahl 2013). As a result, there is an increasingly urgent call for validation and verification of published research results, both within the academic community and the public at large (e.g. Naik 2011; Zimmer 2012; Begley 2012; Editorial 2013a, b; Branco 2012). The discussion surrounding the reliability of published scientific results is often focused on the Life Sciences, especially in the media because of the immediate relevance of work in genomics, neuroscience, and other health-related areas to the public good; but the problem extends to all empirically-based

The original version of this article was revised.

✉ Nancy Ide
ide@cs.vassar.edu

[1] Departamento de Informática, Faculdade de Ciências de Lisboa, Campo Grande 1749-016 Lisboa, Portugal

[2] Biomedical Text Mining Group, University of Colorado School of Medicine, Boulder, CO 80309, USA

[3] Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands

[4] Department of Computer Science, Vassar College, Poughkeepsie, NY, USA

[5] Institute for Computational Linguistics "A. Zampolli", CNR, Pisa, Italy

disciplines, including human language technology (HLT). In fact, several recent articles have reported on reproducibility and/or replication problems in the HLT field (e.g., Johnson et al. 2007; Poprat et al. 2008; Gao and Vogel 2008; Caporaso et al. 2008; Kano et al. 2009; Fokkens et al. 2013; Hagen et al. 2015), and two recent workshops[1] have addressed the need for replication and reproduction of HLT results. However, there is no established venue for publications on the topic, and perhaps more problematically, research that investigates existing methods rather than introducing new ones is often implicitly discouraged in the process of peer review.[2]

To address this need, *Language Resources and Evaluation (LRE)*, the premier journal for publication of papers concerning resources that support HLT research as well as evaluation of both resources and results, is acting to encourage the discussion and advancement of what is commonly referred to as *replicability* and *reproducibility* in the field of Human Language Technology. Researchers have not always used these two terms consistently, as discussed in Liberman (2015); here we adopt the distinction between the two terms put forward in Stodden et al. (2014):

> Replication, the practice of independently implementing scientific experiments to validate specific findings, is the cornerstone of discovering scientific truth. Related to replication is reproducibility, which is the calculation of quantitative scientific results by independent scientist using the original datasets and methods. (*Preface, p. vii*)

It should be noted that despite efforts to distinguish reproducibility and replicability (e.g., by definig "levels" of reproducability Dalle (2012)), the line between the two is not always clear. What is clear is that whether for the purposes of replication or reproduction of prior results, access to the resources, procedures, parameters, and test data used in an original work is critical to the exercise. It has been argued Ince et al. (2012) that insightful reproduction can be an (almost) impossible undertaking without access to the source code, resources (lexica, corpora, tag-sets), explicit test sets (e.g., in case of cross-validation), procedural information (e.g., tokenization rules), and configuration settings, among others[3]; and it has been shown that source code alone is not sufficient to reproduce results Louridas and Gousios (2012). Awareness of the importance of open experiments, in which all required resources and information are provided, is evident in publications in high-profile journals such as *Nature* Ince et al. (2012) and initiatives such as *myExperiment*[4] and gitXiv[5]. However, as discussed in Howison and Herbsleb (2013), even though its importance is increasingly recognized, often not enough (academic) credit is given for making the code and resources used to produce a set of results available.

[1] Workshop on research results reproducibility and resources citation in science and technology of language, Branco et al. (2016), http://4real.di.fc.ul.pt; Replicability and reusability in natural language processing: From data to software sharing, http://nl.ijs.si/rrnlp2015/.

[2] Consider, for example, the question "How novel is the presented approach?" that appears on many HLT conference and journal review forms.

[3] See Mende (2010) for a comprehensive list of the information required to adequately replicate results

[4] http://www.myexperiment.org.

[5] A repository for "open collaborative computer science": http://www.gitxiv.com/

By establishing a special section on Replicability and Reproducibility, *LRE* is encouraging submissions of articles providing positive or negative quantitative assessment of previously published results in the field. We also encourage submission of position papers discussing the procedures for replication and reproduction, including those that may be specific to HLT or could be adopted or adapted from neighboring areas, as well as papers addressing new challenges posed by replication studies themselves. Submissions outlining proposals for solutions to the replicability/reproducibility problem and/or describing platforms that enable and support "slow science"[7] and open, collaborative science in general are also welcome. Articles accepted for publication on the theme will be highlighted in a special section of the *LRE* issue in which they appear, under the heading "replicability and reproducibility". Three members of the *LRE* Editorial Board (António Branco, Kevin Bretonnel Cohen, and Piek Vossen) have been appointed to oversee the reviewing process for submissions addressing the topic.

At the same time, in order to encourage the availability of the resources required for adequate replication and reproduction of research results, the journal is also strongly encouraging the authors of submissions reporting novel research results to provide full and open access to these materials where possible, by including information about where these materials can be obtained (e.g., a github or gitXiv repository, a URL for a Jupyter Notebook[8], etc.).[9] Our review form is being modified to reflect this new emphasis, by asking reviewers if full materials have been made openly available.

LRE accepts full papers, survey articles, and Project Notes. Submissions in any of these categories are appropriate for papers reporting on replication/reproduction experiments as well as papers addressing issues surrounding the topic. *LRE* Project Notes, in particular, provide a venue for publication of information about the availability of materials and experimental data for experiments previously reported in *LRE* or elsewhere, or data that reflect interim results that can be used in replication/reproduction studies and upon which others can profitably build or expand.

LRE's fostering of submissions reporting results of replicability and reproducibility studies and reports on experimental resource availability reflects its commitment to fostering a fundamentally collaborative (rather than competitive) mindset within the field. In addition, by providing a respected venue for publications on the topic, the journal wishes to reiterate its commitment to ensuring adequate academic credit for research and development activities–including both replicability and reproducibility studies and publication of experimental resources–that traditionally have not been well-recognized in HLT.

[7] http://slow-science.org.

[8] http://jupyter.org.

[9] We recognize that a full definition of what is required for replication/reproduction exercises for HLT is not clearly established, and hope it will be addressed in discussions within the community.

References

Anderson, J. E., Aarts, A. A., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Begley, S. (2012). In cancer science, many "discoveries" don't hold up. Reuters. http://www.reuters.com/article/us-science-cancer-idUSBRE82R12P20120328

Bohannon, J. (2013). Who's afraid of peer review. *Science*, *342*(6154), 60–65.

Branco, A. (2012). Reliability and meta-reliability of language resources: Ready to initiate the integrity debate? In *The 12th workshop on treebanks and linguistic theories* (TLT12).

Branco, A., Calzolari, N. & Choukri, K. (Eds.). (2016). *Workshop on research results reproducibility and resources citation in science and technology of language. 10th language resources and evaluation conference (LREC2016).*

Buchert, T. & Nussbaum, L. (2011). Leveraging business workflows in distributed systems research for the orchestration of reproducible and scalable experiments. In *9ème édition de la conférence MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication-MajecSTIC 2012*.

Caporaso, JG., Deshpande, N., Fink, JL., Bourne, PE., Cohen, KB. & Hunter, L. (2008). Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. In *Pacific symposium on biocomputing. Pacific symposium on biocomputing*. NIH Public Access, (p. 640).

Dalle, O. (2012). On reproducibility and traceability of simulations. In *Proceedings of the 2012 winter simulation conference (WSC)* (pp. 1–12). IEEE.

Drummond, C. (2009). Replicability is not reproducibility: nor is itgood science. In *Twenty-Sixth international conference on machinelearning: Workshop on evaluation methods for machine learning IV*.

Editorial. (2013a). Announcement: Reducing our irreproducibility. *Nature News Nature*

Editorial. (2013b). Unreliable research: Trouble at the lab. *The Economist*. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS ONE*, *4*(5), e5738.

Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P. & Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (Vol. 1, pp. 1691–1701).

Gao, Q. & Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software engineering, testing, and quality assurance for natural language processing, association for computational linguistics* (pp. 49–57).

Hagen, M., Potthast, M., Büchner, M. & Stein, B. (2015). Webis: An ensemble for twitter sentiment detection. *Citeseer*.

Howison, J. & Herbsleb, JD. (2013). Sharing the spoils: Incentives and collaboration in scientific software development. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 459–470).

Hvistendahl, M. (2013). China's publication bazaar. *Science*, *342*(6162), 1035–1039.

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485–488.

Ioannidis, J. P. (2011). An epidemic of false claims. *Scientific American*, *304*(6), 16–16.

Johnson, H. L., Cohen, K. B. & Hunter, L. (2007). A fault model for ontology mapping, alignment, and linking systems. In *Pacific symposium on biocomputing. pacific symposium on biocomputing*. NIH Public Access (p. 233).

Kano, Y., Baumgartner, W. A., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L., et al. (2009). U-compare: Share and compare text mining tools with UIMA. *Bioinformatics*, *25*(15), 1997–1998.

Liberman, M. (2015). Replicability vs. reproducibility or is it the other way around? http://languagelog.ldc.upenn.edu/nll/?p=21956, Accessed 31 Oct 2015.

Louridas, P., & Gousios, G. (2012). A note on rigour and replicability. *ACM SIGSOFT Software Engineering Notes*, *37*(5), 1–4.

Mende, T. (2010). Replication of defect prediction studies: Problems, pitfalls and recommendations. In *Proceedings of the 6th international conference on predictive models in software engineering*, ACM.

Naik, G. (2011). Scientists' elusive goal: Reproducing study results. *Wall Street Journal*, *258*(130), A1.

Poprat, M., Beisswanger, E. & Hahn, U. (2008). Building a biowordnet by using wordnet's data formats and wordnet's software infrastructure: a failure story. In *Software engineering, testing, and quality assurance for natural language processing, association for computational linguistics* (pp. 31–39).

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712.

Stodden, V., Leisch, F., Peng, R. D. (Eds.) (2014). *Implementing reproducible research*. CRC Press, http://www.crcpress.com/product/isbn/9781466561595.

Zimmer, C. (2012). A sharp rise in retractions prompts calls for reform. *The New York Times 16*. http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html.