# A Shared Task of a New, Collaborative Type to Foster Reproducibility

Branco, António; Calzolari, Nicoletta; Vossen, Piek; van Noord, Gertjan; van Uytvanck, Dieter; Silva, João; Gomes, Luis; Moreira, André; Elbers, Willem

*Published in:*
Proceedings of The 12th Language Resources and Evaluation Conference

*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 5539–5545
Marseille, 11–16 May 2020


# A Shared Task of a New, Collaborative Type to foster Reproducibility: A first exercise in the area of language science and technology with REPROLANG2020

**António Branco,[1] Nicoletta Calzolari,[2] Piek Vossen,[3] Gertjan van Noord,[4] Dieter Van Uytvank,[5] João Silva,[1] Luís Gomes,[1] André Moreira,[5] Willem Elbers[5]**
[1]University of Lisbon, Department of Informatics, Faculdade de Ciências, Portugal, antonio.branco@di.fc.ul.pt
[2]Istituto di Linguistica Computazionale, CNR, Pisa, glottolo@ilc.cnr.it
[3]Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, piek.vossen@vu.nl
[4]University of Groningen, Groningen, The Netherlands, g.j.m.van.noord@rug.nl
[5]CLARIN ERIC, Utrecht, The Netherlands, firstname@clarin.eu

**Abstract**

In this paper, we introduce a new type of shared task — which is collaborative rather than competitive — designed to support and foster the reproduction of research results. We also describe the first event running such a novel challenge, present the results obtained, discuss the lessons learned and ponder on future undertakings.

**Keywords:** reproduction, replication, natural language processing, computational linguistics, language technology

## 1. Introduction

Scientific knowledge is grounded on falsifiable predictions and thus its credibility and raison d'être rely on the possibility of repeating experiments and getting similar results as originally obtained and reported. In many young scientific areas, including Natural Language Processing (NLP), acknowledgement and promotion of the reproduction of research results need to be increased (Branco, 2013).

To raise awareness of the importance of reproducibility in NLP, we organised a community-wide shared task at LREC2020—The 12th International Conference on Language Resources and Evaluation—, to elicit and motivate the spread of scientific work on reproduction. This initiative builds on the previous pioneer LREC workshops on reproducibility 4REAL2016 (Branco et al., 2016) and 4REAL2018 (Branco et al., 2018). It follows also the initiative of the *Language Resources and Evaluation* journal, with its special section on reproducibility and replicability (Branco et al., 2017).

Shared tasks are an important instrument to stimulate scientific research and to advance the state of the art in many areas and topics in a measurable fashion. They facilitate competition among research teams that seek to resolve a common problem or task with the best possible solution or performance. The proposed task is typically a well-described yet scientifically challenging problem and the submitted solutions by the different teams are evaluated against the same test sets for comparison (kept secret during the development phase).

In this paper, we introduce a new type of shared task—which is collaborative rather than competitive—designed to support and foster the reproduction of research results: "the calculation of quantitative scientific results by independent scientists using the original data sets and methods", (Stodden et al., 2014, Preface, p. vii).

We also describe the first event running such a novel challenge, present the results obtained, discuss the lessons learned and ponder on future undertakings.

The task, called REPROLANG-The Shared Task on the Reproduction of Research Results in Science and Technology of Language, was organized by ELRA-European Language Resources Association—on the occasion of its 25th anniversary—with the technical support of CLARIN-European Research Infrastructure for Language Resources and Technology, and promoted by a Steering Committee presented in Annex I.

The results of this shared task were presented as in a specific session on reproducibility in the main track program of LREC2020 and the papers describing the contributions of the participating teams are published in its Proceedings, after they had been reviewed and selected as described below.

This paper is organized as follows. We first elaborate on the cooperative nature of the challenge, in Section 2. In Section 3, we describe the process and result of selecting the actual tasks, while in Section 4 we explain the procedures for submission and reviewing. The results are described in Section 5 and the lessons learned in Section 6. Finally, we draw conclusions in Section 7.

## 2. A cooperative challenge

This shared task is a new type of challenge: it is partly similar to the usual competitive shared tasks—in the sense that all participants share a common goal; but it is partly different to previous shared tasks—in the sense that its primary focus is on seeking support and confirmation of previous results, rather than on overcoming those previous results with superior ones. Thus instead of a competitive shared task, with each participant struggling for an individual top system that scores as high as possible above a baseline, this is a cooperative shared task, with participants struggling for systems to reproduce as close as possible the results to an original complex research experiment and thus eventually reinforcing the level of reliability on its results by means of their eventually convergent outcomes.

Concomitantly, like with competitive shared tasks, new ideas for improvement and advances beyond the repro-

duced results are expected to sprout from the participation in such a collaborative shared task.

To the best of our knowledge, the REPROLANG challenge was the first instance of this new type of shared task. Through widely disseminated calls for papers, researchers were invited to reproduce the results of a selected set of articles from NLP, which have been offered by the respective authors or with their consent to be used for this shared task (see Section 3. below for the selected tasks).

In addition, we encouraged submissions that report on the replication of the selected tasks with other languages, domains, data sets, models, methods, algorithms, downstream tasks, etc, in addition to the reproduction itself. These submissions may give insight into the robustness of the replicated approaches, their learning curves and potential for incremental performance, their capacity of generalization, their transferability across experimental circumstances and even in real-life scenarios, their suitability to support further progress, etc.

## 3. The tasks

The REPROLANG challenge comprised a number of tasks each consisting in reproducing the experimental results from a previously published paper.

The papers to be reproduced were selected by a Task Selection Committee, presented in Annex II. This committee announced an open call for paper offerings, asking for authors of published papers to offer their paper for reproduction. Authors who offered their paper for reproduction were asked to provide a short motivation indicating the reasons why they believed their paper to be suitable for the reproduction exercise.

In addition, the Task Selection committee contacted authors of specific papers directly, for papers which the committee found particularly promising. In total, 20 potential papers were collected: 12 papers by means of the open call, and 8 further papers that were invited by the selection committee directly. In all cases, authors accepted their papers to be reproduced at REPROLANG.

The Task Selection committee then made a further selection from these 20 papers, aiming at high quality, diversity of domains and approaches, potential of triggering further advances, etc. This resulted in the final list of 11 papers to be included as target papers for reproduction for REPROLANG.

The tasks consisted in reproducing one of those selected papers. Participants were expected to obtain the data and tools for the reproduction from the information provided in the paper. Using the description of the experiment was part of the reproduction exercise. The list of papers was the following:

### Chapter A: Lexical processing

**Task A.1: Cross-lingual word embeddings**

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018. "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings". In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 789–798, (Artetxe et al., 2018).

**Task A.2: Named entity embeddings**

Newman-Griffis, Denis, Albert M Lai, and Eric Fosler-Lussier. 2018. "Jointly Embedding Entities and Text with Distant Supervision". In Proceedings of The Third Workshop on Representation Learning for NLP, pp. 195–206, (Newman-Griffis et al., 2018).

### Chapter B: Sentence processing

**Task B.1: POS tagging**

Bohnet, Bernd, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. "Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings". In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 2642–2652, (Bohnet et al., 2018).

**Task B.2: Sentence semantic relatedness**

Gupta, Amulya, and Zhu Zhang. 2018. "To Attend or not to Attend: A Case Study on Syntactic Structures for Semantic Relatedness". In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 2116–2125, (Gupta and Zhang, 2018).

### Chapter C: Text processing

**Task C.1: Relation extraction and classification**

Rotsztejn, Jonathan, Nora Hollenstein, and Ce Zhang. 2018. "ETH-DS3Lab at SemEval-2018 Task 7: Effectively Combining Recurrent and Convolutional Neural Networks for Relation Classification and Extraction". In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), pp. 689–696, (Rotsztejn et al., 2018).

**Task C.2: Privacy preserving representation**

Li, Yitong, Timothy Baldwin, and Trevor Cohn. 2018. "Towards Robust and Privacy-preserving Text Representations". In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 25–30, (Li et al., 2018).

**Task C.3: Language modelling**

Howard, Jeremy, and Sebastian Ruder. 2018. "Universal Language Model Fine-tuning for Text Classification". In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 328–339, (Howard and Ruder, 2018).

### Chapter D: Applications

**Task D.1: Text simplification**

Nisioi, Sergiu, Sanja Stajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. "Exploring Neural Text Simplification Models". In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), pp. 85–91, (Nisioi et al., 2017).

**Task D.2: Language proficiency scoring**

Vajjala, Sowmya, and Taraka Rama. 2018. "Experiments with Universal CEFR classifications". In Proceedings of Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 147–153, (Vajjala and Rama, 2018).

**Task D.3: Neural machine translation**

Vanmassenhove, Eva, and Andy Way. 2018. "SuperNMT: Neural Machine Translation with Semantic Supersenses

and Syntactic Supertags". In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 67–73, (Vanmassenhove and Way, 2018).

### Chapter E: Language resources

**Task E.1: Parallel corpus construction**

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. "PaCCSS-IT: A Parallel Corpus of Complex-Simple Sentences for Automatic Text Simplification". In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), pp. 351–361, (Brunato et al., 2016).

## 4. The submissions

After selecting the target papers, we published a call for papers through a dedicated section within the LREC2020 website and through various channels, addressing the LREC community. The call for papers explained the procedure and listed the selected papers for reproduction. Submissions had to consist of two parts. On the one hand, an up to eight page length report on the reproduction, documenting how the results of the target paper were reproduced, discussing reproducibility challenges, informing on time, space or data requirements found concerning training and testing, pondering on lessons learned, elaborating on recommendations for best practices, etc. On the other hand, the software used to obtain the results reported in the paper had to be made available as a Docker container[1] through a project in Gitlab.

### 4.1. Reproducing the reproductions

The submitted software was run by the Technical Committee, composed by members from CLARIN ERIC and from the University of Lisbon, presented in Annex IV.
Submissions had to include the following elements:

1. URL address of the gitlab.com project
2. commit hash and tag of the release to be reviewed
3. URL of a tar.gz file containing the datasets
4. MD5 checksum of the above tar.gz

The project in gitlab.com had to be made public within 2 days after the submission deadline for papers. In order to ensure a runtime environment as similar as possible to that of the submitting authors, a framework based on containers was introduced.[2]
Each submission was assessed with respect to the overall quality of the container images. This included assessing the use of best practices such as (but not limited to):

- Using version pinning of software packages
- Not installing dependencies at run time
- Not including large data sets in the container image
- Using tags if git repositories are cloned inside a container image
- Including scripts directly in the container image instead of mounting these from the host
- Triggering all the experiment's scripts from the entrypoint of the container image
- Not uploading container images manually to the container registry

These assessments where included in the reproduction report provided to the reviewers.
By following these instructions, each submission provided a public git repository on gitlab.com which defined a docker container image. In some cases authors ran into size limitations of the gitlab.com service. After interaction with the respective authors, it was agreed that the Technical Committee would review the submissions and build the container images locally from the submitted release tag.
In retrospect, a size limit on gitlab.com was causing issues because there was an optimization missing in the image generation process. By sharing the full output between the build and release stages, the size limit threshold on gitlab.com was triggered. By merging the two stages and no longer requiring the build output to be shared between stages, we were able to avoid hitting this threshold. With this new approach we were able to build all affected container images according to the guidelines.
With all container images available and a well-defined process in place to run the submissions, we started to provision a number of virtual private servers[3] (VPS), of which some had GPU support. While some of the submissions ran without issues, some had obvious errors in the workflow. Still others had subtle, unexpected issues, such as the use of libraries requiring the availability of specific CPU instructions only present in some CPU models. It turned out that one of the VPS instances did not have the appropriate instruction set available. On another occasion an experiment failed due to a lack of GPU memory. Our instances had 8GB of GPU memory, so for this case we provisioned a new instance with 12GB and where able to successfully run the submission.
To check for possible hard-coded results, we proceeded with ablation of the input data set for each experiment that successfully finished before the review deadline.
To reach a reasonable degree of confidence that the target results to be reproduced had not been hard-coded by the authors submitting the software, it was enough to alter the test data in some way and check whether the results also changed. Since we are only concerned with causing a change in the output, ablation consisted of altering the test files by discarding several entries (for instance, taking only the first 100 lines of a test file). Although ablation was generally straightforward, it had to be customized for each submission, as the test file locations and formats were different between submissions. Each experiment with an ablated input data set was run a second time over this data

[1] https://www.docker.com/

[2] Technical details on the framework can be found here: https://gitlab.com/CLARIN-ERIC/reprolang

[3] A virtual private server is a virtual machine provided as a service

set and the respective output was also provided in the report made available to the reviewers.
It should be mentioned that this ablation method does not provide a unequivocal proof that the results had not been hard-coded, rather it offers a good balance between confidence that the results were not hard-coded and the verification effort.

### 4.2. Reviewing the papers

The reviewing of the submitted papers was undertaken by the Program Committee, presented in Annex III, with the help of the authors of the target papers.
Each submission was reviewed by at least 3 reviewers (anonymous to the submitting authors). Additionally, it was commented on by the authors of the target paper being reproduced. These commentaries adhered to the same format as the review reports and were provided to the submitting authors anonymized and side by side with the latter—so, review reports and commentaries could not be told one from the other by the submitting authors.
For the reviewing, we defined a specific form tailored to the task. In addition to the usual criteria such as appropriateness, clarity, soundness, correctness and thoroughness, the reviewers were instructed to consider the reproduction report from the Technical Committee, who tested the submission by reproducing the reproduction. For this, the reviewing form was extended with some additional criteria:

#### Reproduction success

On the basis of the CLARIN reproduction report, reviewers were asked to score the submission for the following scale, addressing the overall success of the reproduction:

- 5 = Reproduction of the results reported by the respective authors completed without any problem. The paper provided enough information.
- 4 = Reproduction completed but some technical difficulties were found and/or the paper did not provide sufficiently detailed information.
- 3 = Reproduction completed but running it was cumbersome for various reasons and/or documentation was not clear.
- 2 = The means to reproduce were provided by the respective authors but reproduction was not completed or reported results were not generated.
- 1 = It was impossible to run or start the reproduction process given the provided materials and/or information.

#### Reproduction score

Separately, the reviewers were asked to assess to what extent the same results were produced:

- 5 = Reproduction of the results reported by the respective authors completed without any problem. The scores obtained match the ones indicated in the paper.
- 4 = Reproduction completed. Although a few scores obtained do not exactly match the ones indicated in the paper these differences are not essential.
- 3 = Reproduction completed but some scores obtained do not match the ones indicated in the paper. One can still say the overall results are roughly aligned with the ones reported by the authors but there are notable differences.
- 2 = Reproduction completed but there are so many scores that do not match the ones indicated in the paper that one cannot really say that they are aligned with the ones reported by the authors.
- 1 = Either it was not possible to run the replication or the scores obtained deviate so much from the originally reported results that they falsify the claims of the submitted paper.

#### Meaningful reflection

Submissions were further assessed for the degree of reflection concerning their level of reproducibility. Does the authors make clear where the problems, if any, sit with respect to reproduction of the paper they addressed for reproduction?

- 5 = Thoughtful reflection about the addressed task. Good job given the space constraints.
- 4 = Mostly solid reflection, but some aspects are lacking or under scrutinized.
- 3 = Reflection is somewhat helpful, but it could be hard for a reader to determine exactly how this work reflects on the task.
- 2 = Only partial awareness and understanding of the task, or a flawed reflection.
- 1 = Little understanding of the task, or lacks necessary reflection.

#### Replication extra-mile

Finally, the reviewers were asked to asses to what extent the reproduction effort included other languages, domains, data sets, models, methods, algorithms, downstream tasks, etc.

- 5 = In addition to reproducing the results, a wide array of replication results are reported that have the potential to substantially help other people's ongoing research.
- 4 = Some replication results are reported that may help other people's ongoing research.
- 3 = Interesting replication exercise though with a limited range.
- 2 = Marginally interesting.
- 1 = There is no replication reported.

## 5. Selection, presentation and publication

We received 18 submissions, of which 11 were retained for detailed reviewing after cursory inspection that filtered out 7 cases of some sort of equivocation and/or gross formal inadequacy to the submission requirements, including the mandatory co-submission of the companion software.

These 11 submissions addressed the reproduction of 7 out of the 11 shared tasks, that is of the 11 papers offered to be reproduced (Section 3), namely: Task A.1 Cross-lingual word embeddings (2 submissions), Task B.1 POS tagging (1), Task C.1 Relation extraction and classification (1), Task C.3 Language modelling (1), Task D.1 Text simplification (1), Task D.2 Language proficiency scoring (4), Task D.3 Neural machine translation (1).

All 11 submissions were accepted for publications and presented as posters at LREC2020 main track. One of the papers was selected as the best paper, namely (Huber and Çöltekin, 2020), and presented in the oral session dedicated to the shared task right after the poster session.

In the oral session, which followed the poster session, the initial presentation of the best paper was followed by a presentation of the current paper, which in turn was followed by a discussion open to all participants in the task and to the audience in view of collecting suggestions for improvements on future editions of REPROLANG — focusing on NLP in particular —, and on the model of the new collaborative shared task — addressing the fostering of reproducibility in general.[4]

## 6. Lessons learned

This was a first exercise in running a new type of collaborative shared task. To assess its viability, there were a number of settings that called for special monitoring. Other aspects appeared also as crucial during the organization of the event. We report on the most important below.

**Participation of authors of reproduced papers.** Inviting the authors of the papers offered to be reproduced (by them) to contribute with commentaries to the submissions was very positive. In general, the authors produced detailed commentaries, in average lengthier than the reports by the reviewers and very much to the point.

As shared tasks like REPROLANG aim at fostering reproducibility, and ultimately to open the way for a research culture where reproduction papers are accepted in the main tracks of conferences, a desirable evolution of the shared task is not to offer a list of pre-selected papers with the consent of their authors, but rather to allow the submissions to be about the reproduction of any paper selected by the submitting authors.

Judging from the reproduction papers submitted to this first edition of REPROLANG, in general, the target papers presented no special problems for their reproduction. It should be taken into account, however, that these authors offered their papers to be reproduced, or accepted the invitation to do so. It is expected that the full usefulness of reproduction exercises will unfold when any paper can be targeted to be reproduced, including those not offered by their authors, and specially those reporting very outstanding results and progress.

It is an open question, what could be the type of involvement of the original authors and how productive their contribution can be in that scenario.

**Partial reproduction or mere competing replication.** We handled submissions that either presented only a partial reproduction of the proposed task, or instead of reproducing it just presented an alternative solution for the problem, as in a submission of the usual, non-reproduction type. For different reasons, these types of submissions are not helpful to assess the reproducibility of the target paper.

In future reproduction shared tasks, it should be made explicit in the call for papers that partial reproductions or alternative resolutions will be rejected for publication.

**Reproducing the reproductions.** As reproduction aims at checking whether research results can be repeated, it makes sense to take care that their reproductions are themselves reproducible in order to ensure that the reproduction is fair and helps to bring more epistemological clarification and certainty rather than more doubt about the original results. As we were careful in reproducing the reproductions (Section 4.1), we reinforced our conviction that this is a very important and positive requirement for collaborative shared tasks. With this reproducing of the reproductions we also learned a number of other lessons.

This exercise represents a heavy burden on the side of the organizers in terms of manpower and computational resources. For this to be kept manageable within the period available between submission and notification of accepted papers, some cap should be set on the time and computational resources needed for reproduction in order for a submission to be accepted for review. This should be announced explicitly in the call for papers.

Additionally, it should be stated explicitly that a reproduction submission whose software outputs errors, runs endlessly or for too long given the resources and time available for evaluation will be rejected whatever the results reported or the quality of the submitted paper by itself. An improvement for forthcoming editions will be to announce available memory, CPUs and GPUs, in the call for papers.

**Moving out of a shared task.** If as said above a commendable goal of collaborative shared tasks is to ultimately open the way for a research culture where reproduction papers are accepted in the main tracks of regular conferences, in such a more open scenario it should not be expected that reproducing the reproductions is practically feasible, at least in the near future.

Naturally, in these conditions the requirements for the acceptance for publication of a reproduction paper are similar to the requirements that in that respect were met by a target paper, which had not to be reproduced to be accepted by its reviewers. They both have to be reviewed — and rejected or accepted — on the basis of what it is reported in them.

One may argue that, if wrong about the target paper — in particular about the eventual claim that the latter is not reproducible —, a reproduction paper can be more damag-

[4]The oral session took place after the camera-ready version of the present paper had to be ready, in time to be included into the proceedings of the conference, and its outcome has to be documented elsewhere.

ing to the reputation of the target authors, and to scientific progress, than a non-reproduction paper that is wrong. That this, however, may not be the case becomes apparent if one takes into account that a non-reproduction paper of the usual kind is accepted because it claims to overcome some state of the art, thus leaving behind in the dust of history the work and results of other authors. If this claim turns out to be wrong, this can be also damaging to these other authors even though their work is not the target of a reproduction paper.

A suggestion for handling reproduction papers in regular conferences, without resorting to the extra, safety and yet costly step of reproducing reproductions, is to adopt the practice of asking the contribution of the authors of the target papers, who are invited to write a note on the paper reproducing their target paper, which can be appended to the respective reproduction paper after being reviewed by the program committee for appropriateness of content and tone.

All pondered, it does seems viable and highly commendable to have reproduction papers accepted and published in the main tracks of conferences in the future, specially if the target paper reports outstanding breakthroughs. In the meantime, to help this cultural and organizational change to happen in the scientific communities, further editions of collaborative shared tasks may be needed.

## 7. Conclusion

In this paper we described the design of a new type of shared task, which is collaborative rather than competitive, to foster the much needed increase of the practice of reproducing scientific results. We also presented a first reproduction challenge of this type, REPROLANG2020, targeting Natural Language Processing, that was part of the LREC2020 conference main track.

The ultimate goal of this initiative is to help foster and shape a new attitude towards the importance of reproduction for the sustainable progress and credibility of the scientific endeavour, that will eventually lead to have reproduction results and papers as first world citizens of scientific work, conferences and publications.

The settings that had to be conceived and prepared, and the lessons learned with the running of this shared task, documented in the present paper, make us believe that this was a very successful event in view of that ultimate objective, and that it may be a good example to be emulated in other organizational contexts and other scientific areas or communities.

## 8. Acknowledgements

We are very grateful to the authors that offered their papers, or accepted our invitation to do so, to be the target of the reproduction tasks. They are listed in Section 3.

The results reported here were partially supported by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

The computing environment for the replications by the technical committee has been kindly provided by EGI and the EOSC-hub H2020 project (grant agreement 777536) with the dedicated support of the CESGA, CESNET-MCC and RECAS-BARI providers.

## 9. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.

Bohnet, B., McDonald, R., Simões, G., Andor, D., Pitler, E., and Maynez, J. (2018). Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia, July. Association for Computational Linguistics.

Branco, A., Calzolari, N., and Choukri, K. (2016). *Proceedings of the 1st Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language (4REAL2016)*. European Language Resources Association, Paris.

Branco, A., Cohen, K. B., Vossen, P., Ide, N., and Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: introducing an lre special section. *Language Resources and Evaluation*, 51(1):1–5.

Branco, A., Calzolari, N., and Choukri, K. (2018). *Proceedings of the 2nd Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language (4REAL2018)*. European Language Resources Association, Paris.

Branco, A. (2013). Reliability and meta-reliability of language resources: Ready to initiate the integrity debate? In *Proceedings of the 12th Workshop on Treebanks and Linguistic Theories (TLT2013)*, pages 27–36, Sofia, Bulgaria. Bulgarian Academy of Sciences.

Brunato, D., Cimino, A., Dell'Orletta, F., and Venturi, G. (2016). PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas, November. Association for Computational Linguistics.

Gupta, A. and Zhang, Z. (2018). To attend or not to attend: A case study on syntactic structures for semantic relatedness. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2116–2125, Melbourne, Australia, July. Association for Computational Linguistics.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Huber, E. and Çöltekin, Ç. (2020). A reproduction and replication study on CEFR classification as part of re-

prolang 2020. In *12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Li, Y., Baldwin, T., and Cohn, T. (2018). Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia, July. Association for Computational Linguistics.

Newman-Griffis, D., Lai, A. M., and Fosler-Lussier, E. (2018). Jointly embedding entities and text with distant supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206, Melbourne, Australia, July. Association for Computational Linguistics.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada, July. Association for Computational Linguistics.

Rotsztejn, J., Hollenstein, N., and Zhang, C. (2018). ETH-DS3Lab at SemEval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 689–696, New Orleans, Louisiana, June. Association for Computational Linguistics.

Stodden, V., Leisch, F., and Peng, R. D. (2014). *Implementing Reproducible Research*. CRC Press.

Vajjala, S. and Rama, T. (2018). Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana, June. Association for Computational Linguistics.

Vanmassenhove, E. and Way, A. (2018). SuperNMT: Neural machine translation with semantic supersenses and syntactic supertags. In *Proceedings of ACL 2018, Student Research Workshop*, pages 67–73, Melbourne, Australia, July. Association for Computational Linguistics.

## I Steering committee

António Branco, University of Lisbon (chair)
Nicoletta Calzolari, ILC, Pisa (co-chair)
Gertjan van Noord, University of Groningen (chair of Task Selection Committee)
Piek Vossen, VU University Amsterdam (chair of Program Committee)
Khalid Choukri, ELRA/ELDA

## II Task selection committee

Gertjan van Noord, University of Groningen (chair)
Tim Baldwin, University of Melbourne
António Branco, University of Lisbon
Nicoletta Calzolari, ILC, Pisa
Çağrı Çöltekin, University of Tuebingen
Nancy Ide, Vassar College, New York
Malvina Nissim, University of Groningen
Stephan Oepen, University of Oslo
Barbara Plank, University of Copenhagen
Piek Vossen, VU University Amsterdam
Dan Zeman, Prague University

## III Program committee

Piek Vossen, VU University Amsterdam (chair)
Gilles Adda, LIMSI-CNRS, Paris
Eneko Agirre, Basque University
Francis Bond, NanyangTechnical University, Singapore
António Branco, University of Lisbon
Nicoletta Calzolari, ILC, Pisa
Khalid Choukri, ELRA/ELDA
Kevin Cohen, University of Colorado Boulder
Thierry Declerck, DFKI Saarbruecken
Nancy Ide , Vassar College, New York
Antske Fokkens VU University Amsterdam
Karën Fort, University of Paris-Sorbonne
Cyril Grouin, LIMSI-CNRS
Mark Liberman, University of Pennsylvania
John McCrae, Galway University
Margo Mieskes, University of Applied Sciences Darmstadt
Aurélie Névéol, LIMSI-CNRS
Gertjan van Noord, University of Groningen
Stephan Oepen, University of Oslo
Ted Pedersen, University of Minnesota
Senja Pollak, Jozef Stefan Institute, Ljubljana
Paul Rayson, Lancaster University
Martijn Wieling, University of Groningen

## IV Technical committee

Dieter Van Uytvanck, CLARIN (chair)
André Moreira, CLARIN
Twan Goosen, CLARIN
João Ricardo Silva, CLARIN and University of Lisbon
Luís Gomes, CLARIN and University of Lisbon
Willem Elbers, CLARIN