R^3 – *Responsible Replicable Research*

Margot Mieskes

University of Applied Sciences, Darmstadt

June 27, 2025

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmd
FACHBEREICH MEDIA

Are you a researcher

Are you a researcher

What defines a researcher?

Are you a researcher

What defines research?

Research

- creative
- systematic
- control bias
- control error

`https://en.wikipedia.org/wiki/Research`

Research

- creative
- systematic
- control bias
- control error

replicate elements of prior projects

`https://en.wikipedia.org/wiki/Research`

Reproducible vs. Replicable

What does Replicable/Reproducible mean?

Definitions – kind of

Replicability or **repeatability** is a property of an *experiment*: the ability to repeat – or not – the experiment described in a study.

Cohen et al LREC 2018 `https://aclanthology.org/L18-1025/`

Definitions – kind of

Reproducibility is a property of the *outcomes* of an experiment: arriving – or not – at the same conclusions, findings, or values.

Cohen et al LREC 2018 `https://aclanthology.org/L18-1025/`

Dimensions of Reproducibility

Conclusion: Broad induction that is made based on the results of the reported research.

Finding: Relationship between the values for some reported figure of merit with respect to two or more dependent variable.

Value: A number, whether measured (e.g. acount of false positives) or calculated (e.g. a standard deviation).

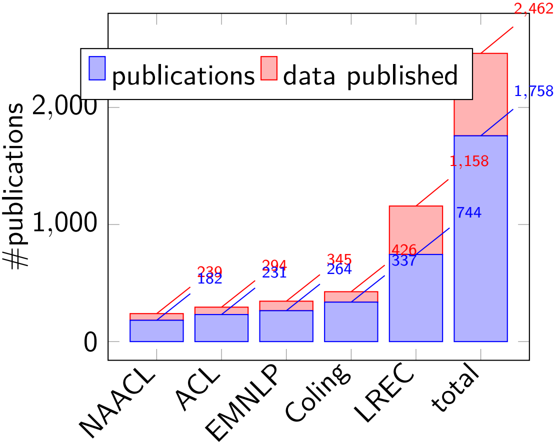
Cohen et al LREC 2018 `https://aclanthology.org/L18-1025/`

Research Artifacts

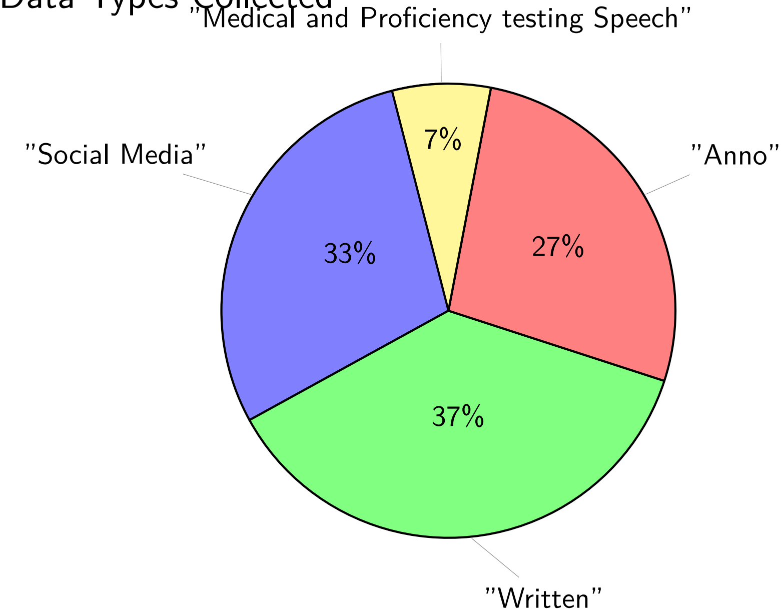
- Data
- Code
- Meta Information

`https://aclanthology.org/W17-1603/`

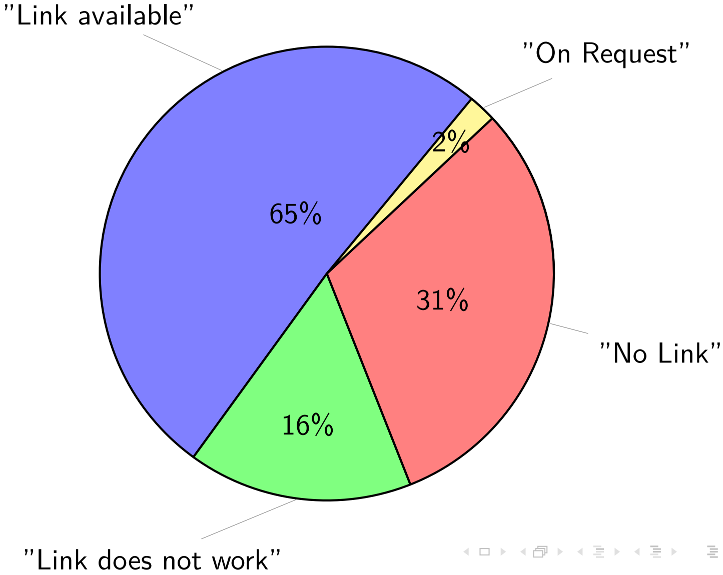
Data Publication



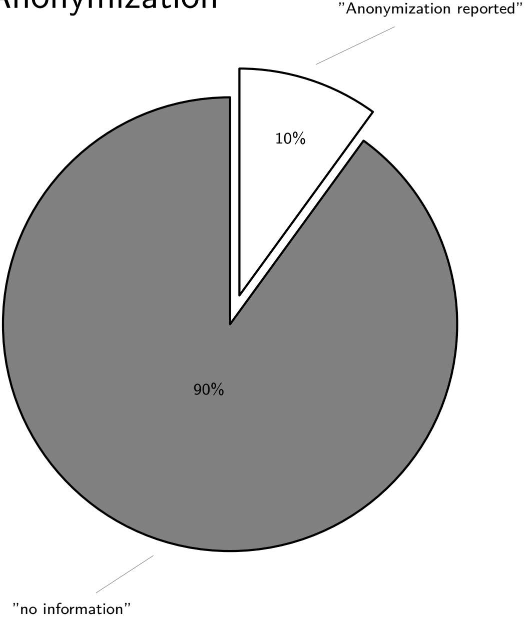
Data Types Collected



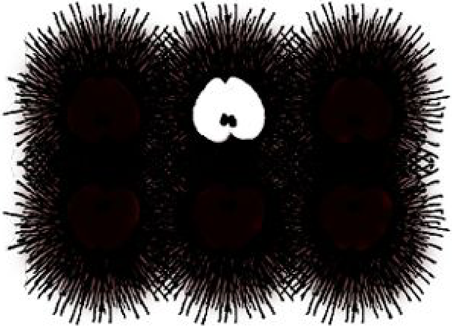
How can I get the data – if at all



Anonymization

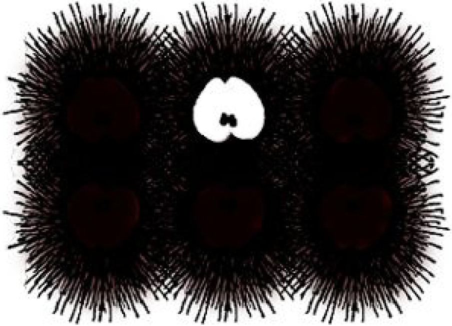


Have we evolved?

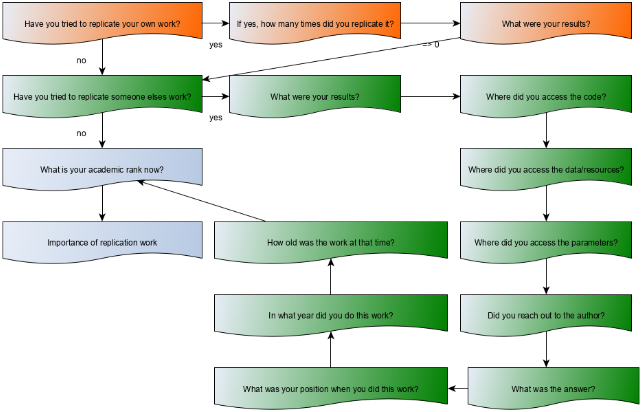


Community

Is it actually a problem?



What did we do?



Have you tried to replicate your own work?

yes

If yes, how many times did you replicate it?

What were your results?

=> 0

no

Have you tried to replicate someone elses work?

yes

What were your results?

Where did you access the code?

no

What is your academic rank now?

Where did you access the data/resources?

Importance of replication work

How old was the work at that time?

Where did you access the parameters?

In what year did you do this work?

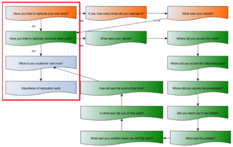
Did you reach out to the author?

What was your position when you did this work?

What was the answer?

`https://aclanthology.org/R19-1089/`

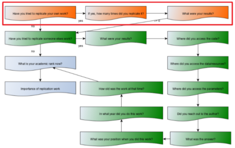
What did we do?


Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
=> 0
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

- Have you tried to replicate your own work?
- Have you tried to replicate someone elses work?
- What is your academic rank?
- How important do you rate replication?

https://aclanthology.org/R19-1089/

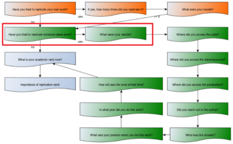
What did we do?


Have you tried to replicate your own work?
yes
If yes, how many times did you replicate it?
What were your results?
=> 0
no
Have you tried to replicate someone elses work?
yes
What were your results?
Where did you access the code?
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

- Have you tried to replicate your own work?
- If yes, how many times did you replicate it?
- What were the results?

https://aclanthology.org/R19-1089/

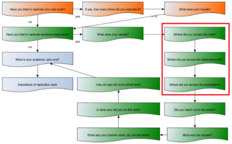
What did we do?


Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
=> 0
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

- Have you tried to replicate someone elses work?
- What were the results?

https://aclanthology.org/R19-1089/

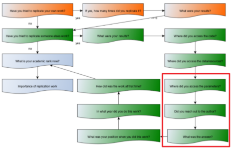
What did we do?


Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
Where did you access the data/resources?
What is your academic rank now?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

- Where did you access the code?
- Where did you access the data/resources?
- Where did you access the parameters?

https://aclanthology.org/R19-1089/

What did we do?


Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

- Did you reach out to the author?
- What was the answer?

https://aclanthology.org/R19-1089/

General Stance towards Replication

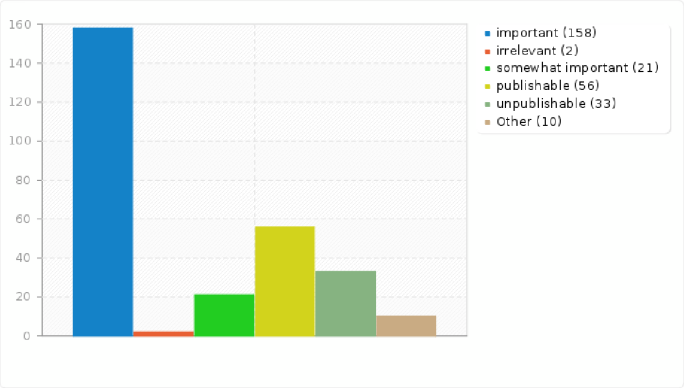


Figure: Importance of Replication.

What are the results?

Part I: Who took part?

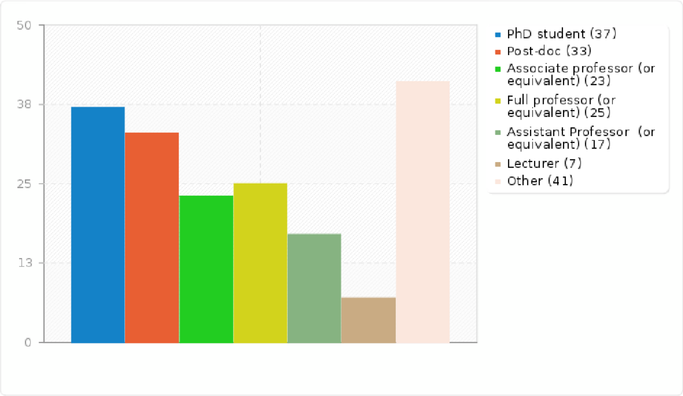
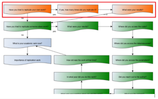


Figure: Seniority of Participants.

What are the results?



Part II: Replicating One's Own Work

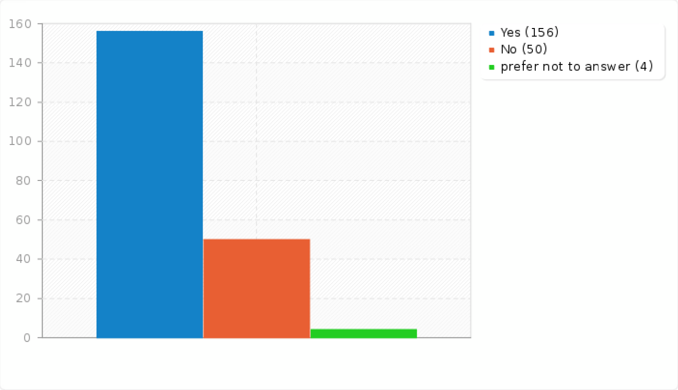
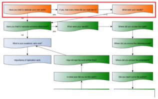


Figure: Replicating one's Own Work.

What are the results?



And its results:

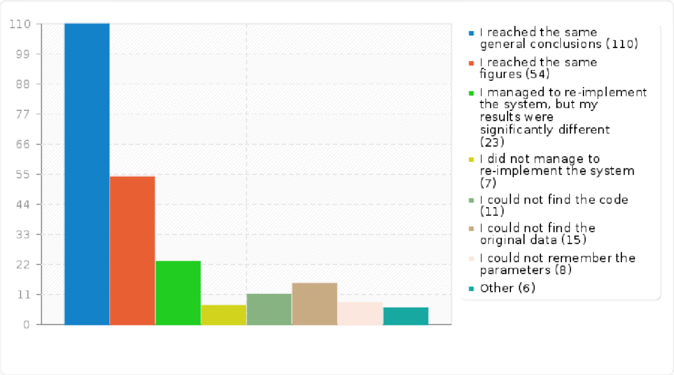


Figure: Results for Replicating one's Own Work.

What are the results?

Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
was the answer?

Part II: Replicating Other Researcher's Work

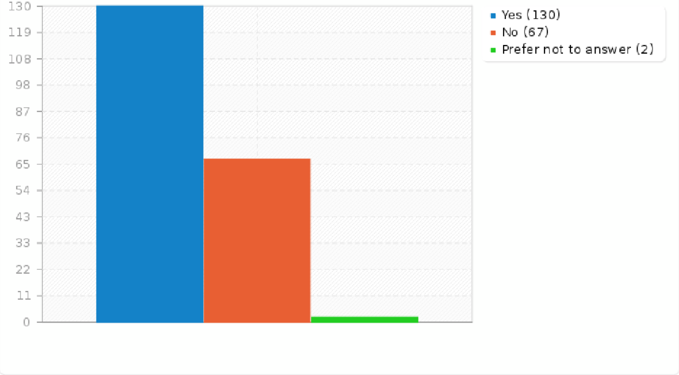


Figure: Replicating Other Researcher's Work.

What are the results?

Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
=> 0
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
Did you reach out to the author?
was the answer?

And its results:

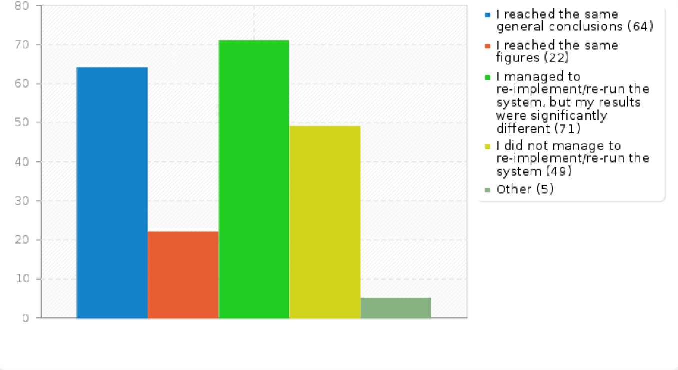
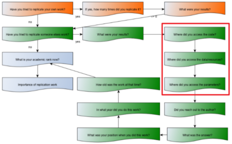


Figure: Results for Replicating Other Researcher's Work.

What are the results?


Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
=> 0
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

Part III: Where are the research artefacts sto

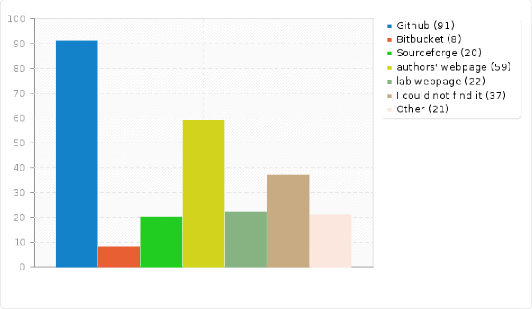


Figure: Where to find the Code.

What are the results?

Part III: Where are the research artefacts sto

Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
=> 0
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

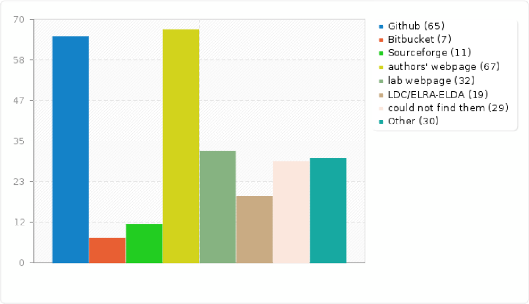
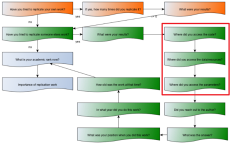


Figure: Where to find the Data.

What are the results?


Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

Part III: Where are the research artefacts sto

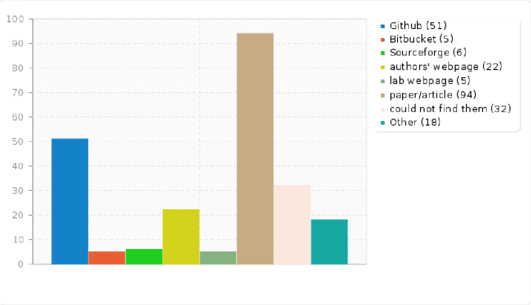


Figure: Where to find the Parameters.

What are the results?

Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

Part III: If all fails, contact the author

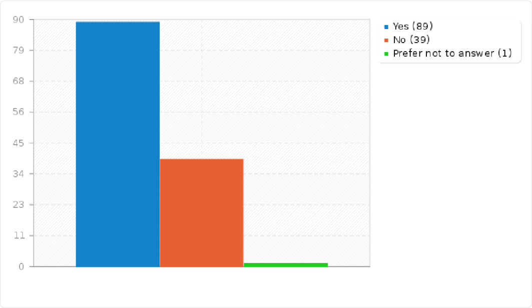


Figure: Contacting the Author.

What are the results?

Have you tried to replicate your own work?
If yes, how many times did you replicate it?
What were your results?
yes
no
Have you tried to replicate someone elses work?
What were your results?
Where did you access the code?
yes
no
What is your academic rank now?
Where did you access the data/resources?
Importance of replication work
How old was the work at that time?
Where did you access the parameters?
In what year did you do this work?
Did you reach out to the author?
What was your position when you did this work?
What was the answer?

Part III: If all fails, contact the author – and

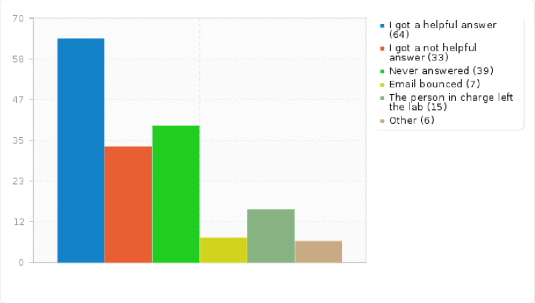


Figure: Types of Answers received from Authors.

What are our conclusions?

- Replicating one's own work is successful in 23.1% of the answers received
- Replicating others work is successful in 40% of the reported cases – in terms of general conclusions
- Research Artefacts are stored in various places – not all of them with any long-term availability
- Research Artefacts are handed over on a person-to-person basis
- Authors are not generally helpful in answering questions – if they answer at all

:-((((

Replicating and Reporting Values

What is actually happening?

`https://aclanthology.org/2022.insights-1.23/`

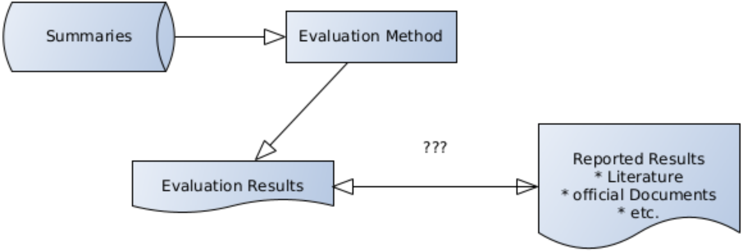
Replicating and Reporting Values

What is actually happening?

A use-case from automatic summarization.....

`https://aclanthology.org/2022.insights-1.23/`

Background


Summaries
Evaluation Method
Evaluation Results
???
Reported Results
* Literature
* official Documents
* etc.

Literature – DUC 2002

Loret	Barrera	Mihalcea	official
S28	S28	S27	S19
S19*	S21*	S28	S28
–	S29*	S21	S21
–	S23*	S29*	S31

Table: * did not beat the baseline according to the source paper.

Results – DUC 2002

Citation	S28	S21	S19
Mihalcea(2004)	0.4703	0.4683	na
Mihalcea(2005)	0.4890	0.4869	na
Loret(2010)	0.4278	0.4149	0.4082
Barrera(2011)	0.4781	0.4754	0.4552

Literature – DUC 2004

Citation	ROUGE-1
Original	0.38224
Yih Wen-Tau (2007)†	0.305
Alguliev (2012)b	0.3822
Ryang(2012)	0.3827
Manna(2012)†	0.3913
Rioux(2014)†	0.3828
Ren(2016)†	0.3788
Wang(2017)†	0.3762

Table: † indicates that parameters have been reported in the publication.

MEAD – DUC 2004

Citation	Result
Erkan(2004)a (added features)	0.38304
Erkan(2004)	0.3758
Alguliev(2012)	0.3673
Hong(2014)†	0.3641
re-run	0.3494

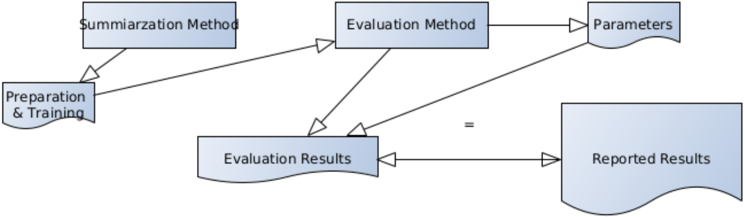
Table: † indicates that parameters have been reported in the publication.

SVM

Sipos(2012)	re-train & eval (95% CI)
0.4066	0.3995 (0.3883–0.4117)

Table: Re-evaluation on DUC 2004 data.

Take home


Summiarzation Method
Evaluation Method
Parameters
Preparation & Training
=
Evaluation Results
Reported Results

Replicating Results

Can we at least draw conclusions?

`https://aclanthology.org/2023.humeval-1.11/`

Replicating Results

Can we at least draw conclusions?

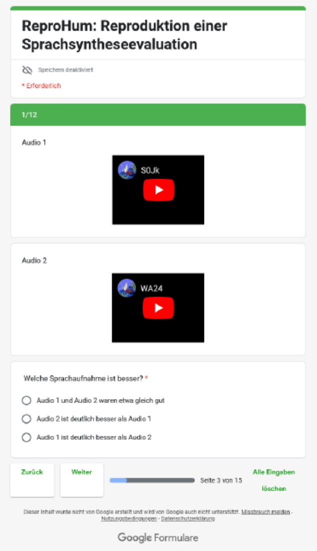
An example from the ReproHum Activities

`https://aclanthology.org/2023.humeval-1.11/`

The Original Study

- Presented by Lux and Vu in 2022
- Creating a TTS System with few resources (data & training time)
- Large Multi-Lingual Model
- Fine-Tuning towards target language
- Models
  - Tacotron2
  - FastSpeech2
- Data
  - English, Greek, Spanish, Finnish, Russian, Hungarian, Dutch, French, German
- Target
  - German
  - 30 Minutes of data
  - Training for 2h
- Comparison: 29 hours of German Data exclusively

Reproducing the Human Evaluation



ReproHum: Reproduktion einer Sprachsyntheseevaluation

Speichern deaktiviert

* Erforderlich

1/12

Audio 1

Audio 2

Welche Sprachaufnahme ist besser? *

○ Audio 1 und Audio 2 waren etwa gleich gut

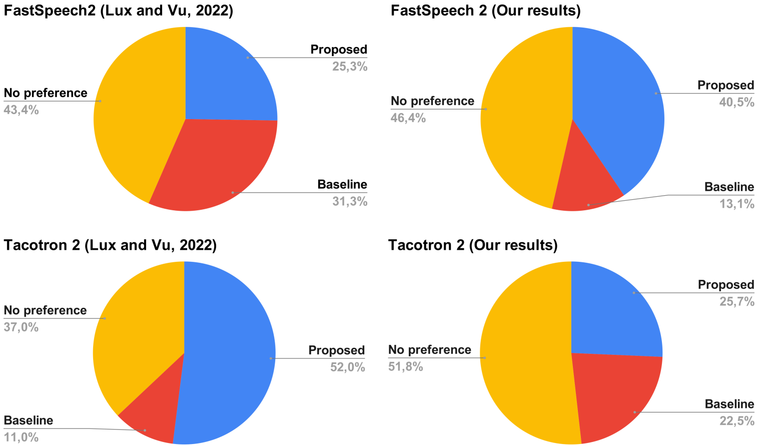
○ Audio 2 ist deutlich besser als Audio 1

○ Audio 1 ist deutlich besser als Audio 2

Dieser Inhalt wurde nicht von Google erstellt und wird von Google auch nicht unterstützt. Missbrauch melden - Nutzungsbedingungen - Datenschutzerklärung

Results

Comparison


FastSpeech2 (Lux and Vu, 2022)

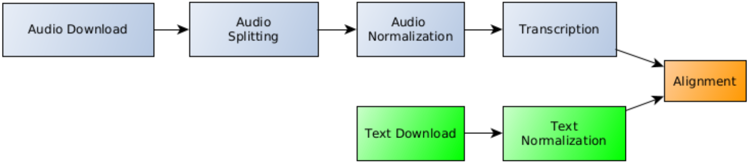
FastSpeech 2 (Our results)

Tacotron 2 (Lux and Vu, 2022)

Tacotron 2 (Our results)

Reproducing the Technical Background – The Voice

Pipeline for creating the Audio-Transcript Data according to Puchtler et al (2021)



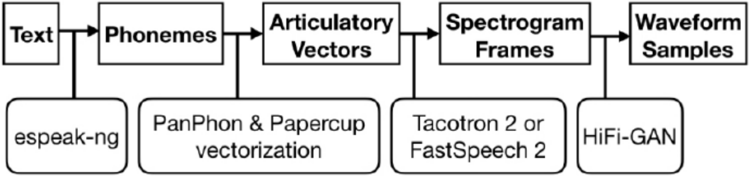
Results

- Dead Links to Speech Model – fixed by original authors
- Hard-coded links to Textual Data dead – fixable
- Missing packages – fixable
- Faulty files from one speaker – not recoverable

Aborted Replication Attempt

Reproducing the Technical Background – The Model

Pipline to create the TTS model according to Lux and Vu (2022)



Results

Modell	HW	Dur Prepr	It	T/It	Tot Dur
Tacotron2 Low Resource	GPU	1:13 min	10,020	1.25 It/sec	2:25 hrs
Tacotron2 full	GPU	50:32 min	100,224	1.4 It/sec	19:54 hrs
Tacotron2 Low Resource	CPU	NA	925	22 sec/It	6 hrs
FastSpeech2 Low Resource	GPU	NA	100,071	4.4 It/sec	6:27 hrs

- Missing packages – fixed
- FFMPEG issues – fixed
- Two versions of the corpus – fixed?
- Number of Training Iterations – fixed?

Results

But

Notable Differences in

- Sound Quality
- Pronounciation
- Prosody

Discussion

Reproduction	Reproducibility	Remarks
Data set	Reproduction had to be abandoned	Mirrors unavailable, software issues
TTS Model	Partially, conclusions were reproduced	Different results, conclusion can be supported
Human Evaluation	Values and results not reproducible	Overall conclusion reproducible

The Verdict

- Replication Crisis
- Failure to Replicate Results even in Computer Science
- Failure to find Necessary Research Artefacts in the NLP domain
- Negative Results when Replicating

Responsible

So how responsible are you taking your tasks as a researcher?

Responsible

So how responsible are you taking your tasks as a researcher?

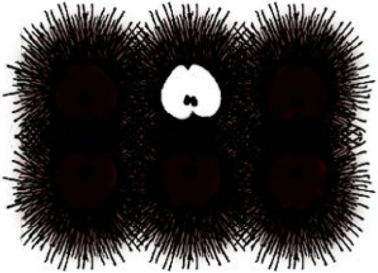
- creative
- systematic
- control
- replicate

Responsible

So how responsible are you taking your tasks as a researcher?

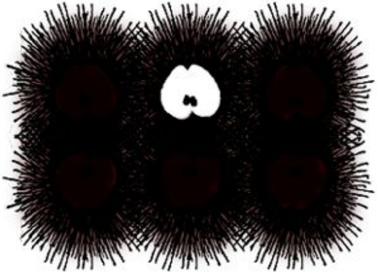
https://de.wikipedia.org/wiki/Datei:Checkliste.svg

Checklist or Guidelines?



Questions?

Many thanks to my students Jacob Benz, Jonathan Baum, Christian Stute,
my colleagues Karen Fort, Aurelie Neveol, Kevin B. Cohen,



and thank you for your attention!