# Reproducibility in Natural Language Processing: what can we learn from NLP?

Simone Paolo Ponzetto

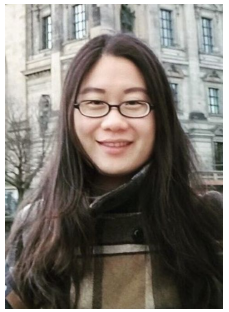
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

# Hi there!

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Professor of Information Systems in Mannheim
- Head honcho of the **NLP and IR group**





# Hi there!

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Professor of Information Systems in Mannheim
- Head honcho of the **NLP and IR group**
- We are part of the larger **Data and Web Science** fleet @ Uni Mannheim

# Today on the menu!

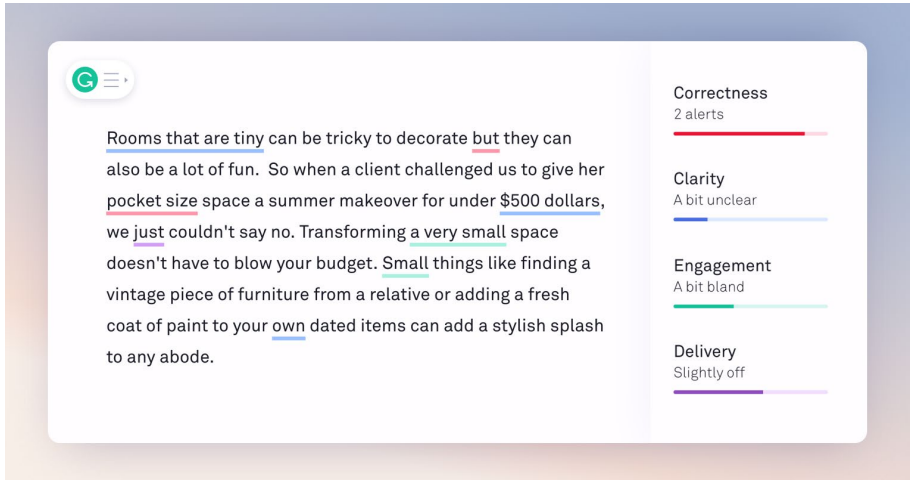
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

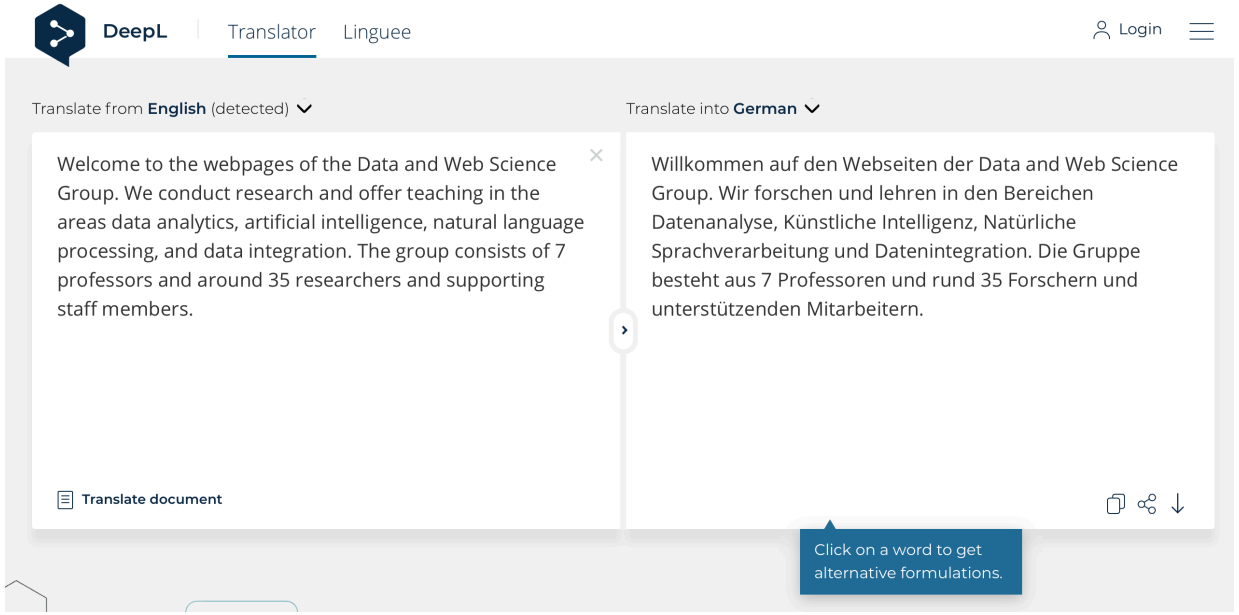
- Natural Language Processing?
- Reproducibility in NLP
  - Community-wide evaluation exercises (a.k.a. “shared tasks”)
  - Open publication models
  - Open data & software
- What can we learn from this?

# Natural Language Processing: some initial thoughts…

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Methods to **automatically process (i.e., understand and generate) natural language data**



# A few applications (from our everyday's life!)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Spelling correction

Text completion

Grammar checking

Speech-to-text and vice versa

Dialogue systems

Question Answering

Summarization

Machine translation

# Evaluation of NLP systems

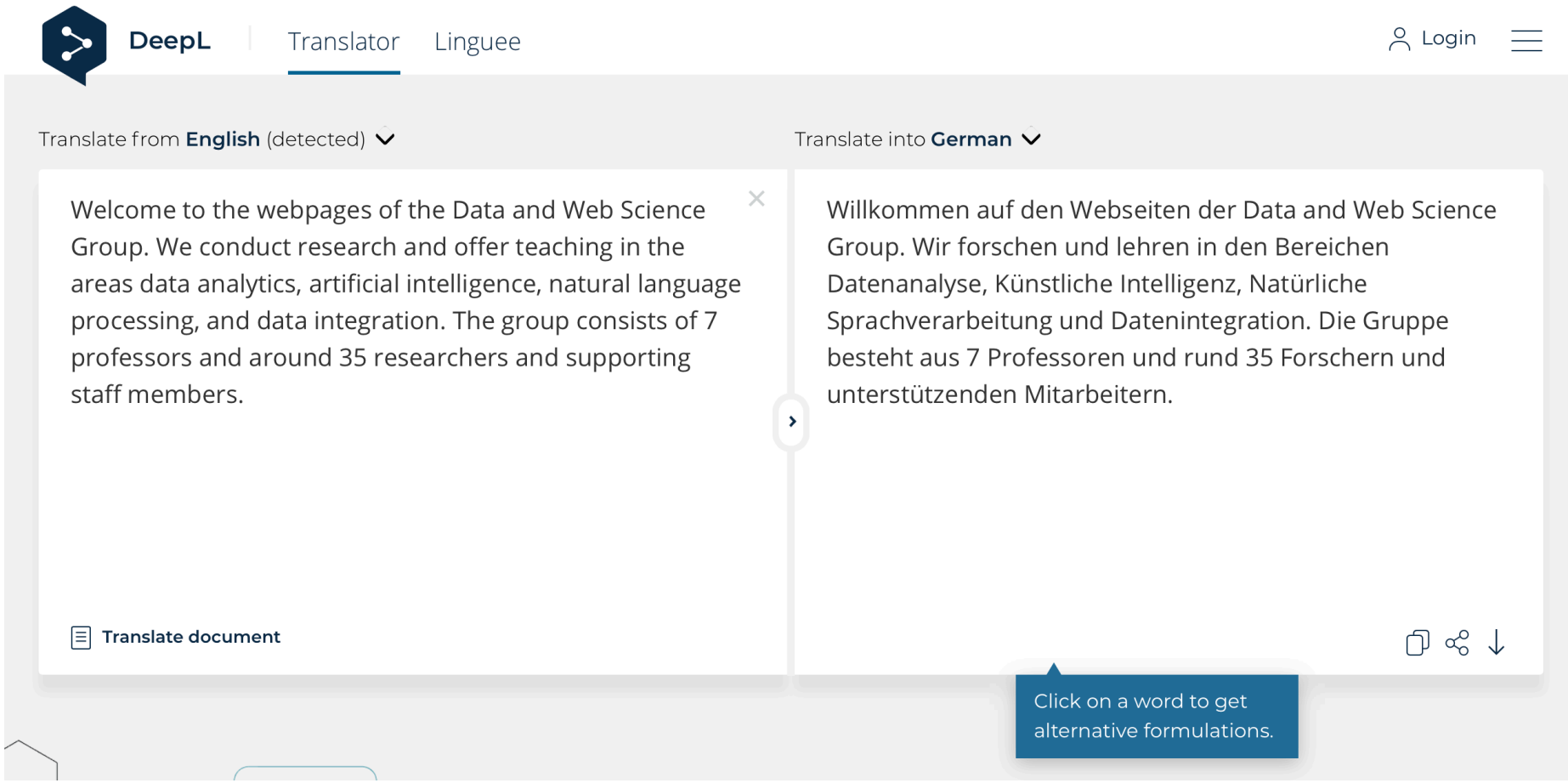
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- The goal of NLP evaluation is to measure one or more *aspects* of an algorithm or a system. Examples:
  - **Performance** (how *good* it is)
  - **Efficiency** (e.g., speed, resource requirements)

- To **establish performance** and **comparability of different systems** we need to define a *experimental setting*
  - Evaluation metrics
  - Evaluation data (language data resources)

# Example: machine translation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



# Evaluating Machine Translation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

| English | Germa |
|---|---|
| Diverging opinions about planned tax reform | Unterschi geplanten |
| The discussion around the envisaged major tax reform continues . | Die Diskussion um die vorgesehene grosse Steuerreform dauert an . |
| The FDP economics expert , Graf Lambsdorff , today came out in favor of advancing the enactment of significant parts of the overhaul , currently planned for 1999 . | Der FDP - Wirtschaftsexperte Graf Lambsdorff sprach sich heute dafuer aus , wesentliche Teile der fuer 1999 geplanten Reform vorzuziehen . |

**Reference translation 1:**
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

**Reference translation 2:**
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

**Machine translation:**
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

**Reference translation 3:**
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

**Reference translation 4:**
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

# Shared tasks?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- In a nutshell, **a competition-like, community-wide evaluation exercise**

- **Shared experimental setting**
  - Data
  - Metrics
- Open for virtually anyone to
  - **Participate**
  - **Describe approach** into a report (typically a workshop paper)
  - **Present** in front of a large audience of experts and fellow colleagues

# Shared task in 2020s

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Data Science competitions are all the rage these days: see Kaggle
- NLP has a long tradition of organising competitions of this kind
- First attempts with MUC / TREC (through funding and support from NIST and DARPA)


Source: Wikipedia

# SemEval

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

View on GitHub

## SemEval-2023

The 17th International Workshop on Semantic Evaluation

**We are pleased to announce the following tasks for SemEval-2023!**

## TASKS

Websites and contact information for individual tasks are given below.

### Semantic Structure

- **Task 1: V-WSD: Visual Word Sense Disambiguation** ([contact organizers], [join task mailing list])
  Alessandro Raganato, Iacer Calixto, Jose Camacho-Collados, Asahi Ushio, Mohammad Taher Pilehvar
- **Task 2: Multilingual Complex Named Entity Recognition (MultiCoNER 2)** ([contact organizers], [join task mailing list])
  Shervin Malmasi, Besnik Fetahu, Sudipta Kar

### Discourse and Argumentation

- **Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup** ([contact organizers], [join task mailing list])
  Giovanni Da San Martino, Jakub Piskorski, Nicolas Stefanovitch, Preslav Nakov
- **Task 4: ValueEval: Identification of Human Values behind Arguments** ([contact organizers], [join task mailing list])
  Johannes Kiesel, Milad Alshomary, Henning Wachsmuth, Benno Stein
- **Task 5: Clickbait Spoiling** ([contact organizers], [join task mailing list])
  Maik Fröbe, Tim Gollub, Matthias Hagen, Martin Potthast
- **Task 6: LegalEval: Understanding Legal Texts** ([contact organizers], [join task mailing list])
  Prathamesh Ashok Kalamkar, Saurabh Kumar Karn, Sachin Malhan, Vivek Raghavan, Shouvik Kumar Guha, Ashutosh Modi

### Medical Applications

- **Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data** ([contact organizers], [join task mailing list])
  Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, Andre Freitas
- **Task 8: Causal medical claim identification and related PICO frame extraction from social media posts** ([contact organizers] [join task mailing list])
  (temporarily anonymous)

### Social Attitudes

- **Task 9: Multilingual Tweet Intimacy Analysis** ([contact organizers], [join task mailing list])
  Jiaxin Pei, Francesco Barbieri, Vítor Silva, Maarten Bos, Yozen Liu, Leor David Jurgens
- **Task 10: Towards Explainable Detection of Online Sexism** ([contact organizers], [join task mailing list])
  Hannah Rose Kirk, Wenjie Yin, Paul Röttger, Bertie Vidgen
- **Task 11: Learning with Disagreements (Le-Wi-Di), 2nd edition** ([contact organizers], [join task mailing list])
  Elisa Leonardelli, Valerio Basile, Gavin Abercombie, Tommaso Fornaciar Plank, Massimo Poesio, Verena Rieser, Alexandra Uma
- **Task 12: AfriSenti-SemEval: Sentiment Analysis for Low-resource A Languages using Twitter Dataset** ([contact organizers], [join task mailin
  Shamsuddeen Hassan Muhammad, Seid Muhie Yimam, Idris Abdulmum Said Ahmad, Saif Mohammad, David Ifeoluwa Adelani, Sebastian Ruder Ousidhoum, Vukosi Marivate, Abinew Ali Ayele, Meriem Beloucif

https://semeval.github.io

# CoNLL

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Conference on Natural Language Learning is a yearly meeting of Special Interest Group on Nature Language Learning (SIGNLL) of the Association for Computational Linguistics

- Since 1999, CoNLL has included a **shared task** in which *training and test data is provided by the organizers* which **allows participating systems to be evaluated and compared in a systematic way**

# CoNLL: previous shared tasks

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

CoNLL

The SIGNLL Conference on Computational Natural Language Learning

## Previous shared tasks

| | | | |
|---|---|---|---|
| 2019 | Cross-Framework Meaning Representation Parsing | English | Proceedings |
| 2018 | Universal Morphological Reinflection | multilingual | Proceedings |
| 2018 | Multilingual Parsing from Raw Text to Universal Dependencies | multilingual | Proceedings |
| 2017 | Multilingual Parsing from Raw Text to Universal Dependencies | multilingual | Proceedings |
| 2017 | Universal Morphological Reinflection | multilingual | Proceedings |
| 2016 | Multilingual Shallow Discourse Parsing | English, Chinese | Proceedings |
| 2015 | Shallow Discourse Parsing | English | Proceedings |
| 2014 | Grammatical Error Correction | English | Proceedings |
| 2013 | Grammatical Error Correction | English | Proceedings |
| 2012 | Modelling Multilingual Unrestricted Coreference in OntoNotes | English, Chinese, Arabic | Proceedings |
| 2011 | Modelling Unrestricted Coreference in OntoNotes | English | Proceedings |
| 2010 | Hedge Detection | English | Proceedings |
| 2009 | Syntactic and Semantic Dependencies in Multiple Languages | multilingual | Proceedings |
| 2008 | Joint Parsing of Syntactic and Semantic Dependencies | English | Proceedings |
| 2007 | Dependency Parsing: Multilingual & Domain Adaptation | multilingual | |
| 2006 | Multi-Lingual Dependency Parsing | multilingual | |
| 2005 | Semantic Role Labeling | English | |
| 2004 | Semantic Role Labeling | English | |
| 2003 | Language-Independent Named Entity Recognition | English, German | |
| 2002 | Language-Independent Named Entity Recognition | Spanish, Dutch | |
| 2001 | Clause Identification | English | |
| 2000 | Chunking | English | |
| 1999 | NP Bracketing | English | |

https://www.conll.org/previous-tasks

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

# GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

**Alex Wang**[1], **Amanpreet Singh**[1], **Julian Michael**[1]
**Omer Levy**[2] **& Samuel R. Bowman**[1]
[1]Cour
[2]Paul
[3]Deep
{ale
{jul
feli

GLUE SuperGLUE Pap

| Rank | Name | Model |
|---|---|---|
| 1 | JDExplore d-team | Vega v1 |
| 2 | Microsoft Alexander v-team | Turing NLR v5 |
| 3 | DIRL Team | DeBERTa + CLEVER |
| 4 | ERNIE Team - Baidu | ERNIE |
| 5 | AliceMind & DIRL | StructBERT + CLEVER |
| 6 | DeBERTa Team - Microsoft | DeBERTa / TuringNLRv4 |

Search or jump to... / Pull requests Issues Marketplace Explore

nyu-mll / GLUE-baselines Public

Watch 29 Fork 152 Star 557

Code Issues 14 Pull requests 1 Actions Projects Security Insights

master 2 branches 0 tags Go to file Add file Code

**w4ngatang** Update data download script; don't download MRPC I... b1c8239 on 10 Dec 2020 141 commits

| | | |
|---|---|---|
| src | Switch to LR 1e-4 | 4 years ago |
| .gitignore | Fix bug with handling pair tasks; add gitignore | 5 years ago |
| README.md | Update paper link. | 4 years ago |
| download_glue_data.py | Update data download script; don't download MRPC IDs or S... | 2 years ago |
| environment.yml | Upgrade to Pytorch 0.4, AllenNLP 0.4; runs but breaks check... | 4 years ago |

README.md

## GLUE Baselines

This repo contains the code for baselines for the Generalized Language Understanding Evaluation (GLUE) benchmark. See our paper for more details about GLUE or the baselines.

**About**

[DEPRECATED] Repo for exploring multi-task learning approaches to learning sentence representations

gluebenchmark.com

Readme
557 stars
29 watching
152 forks

**Releases**

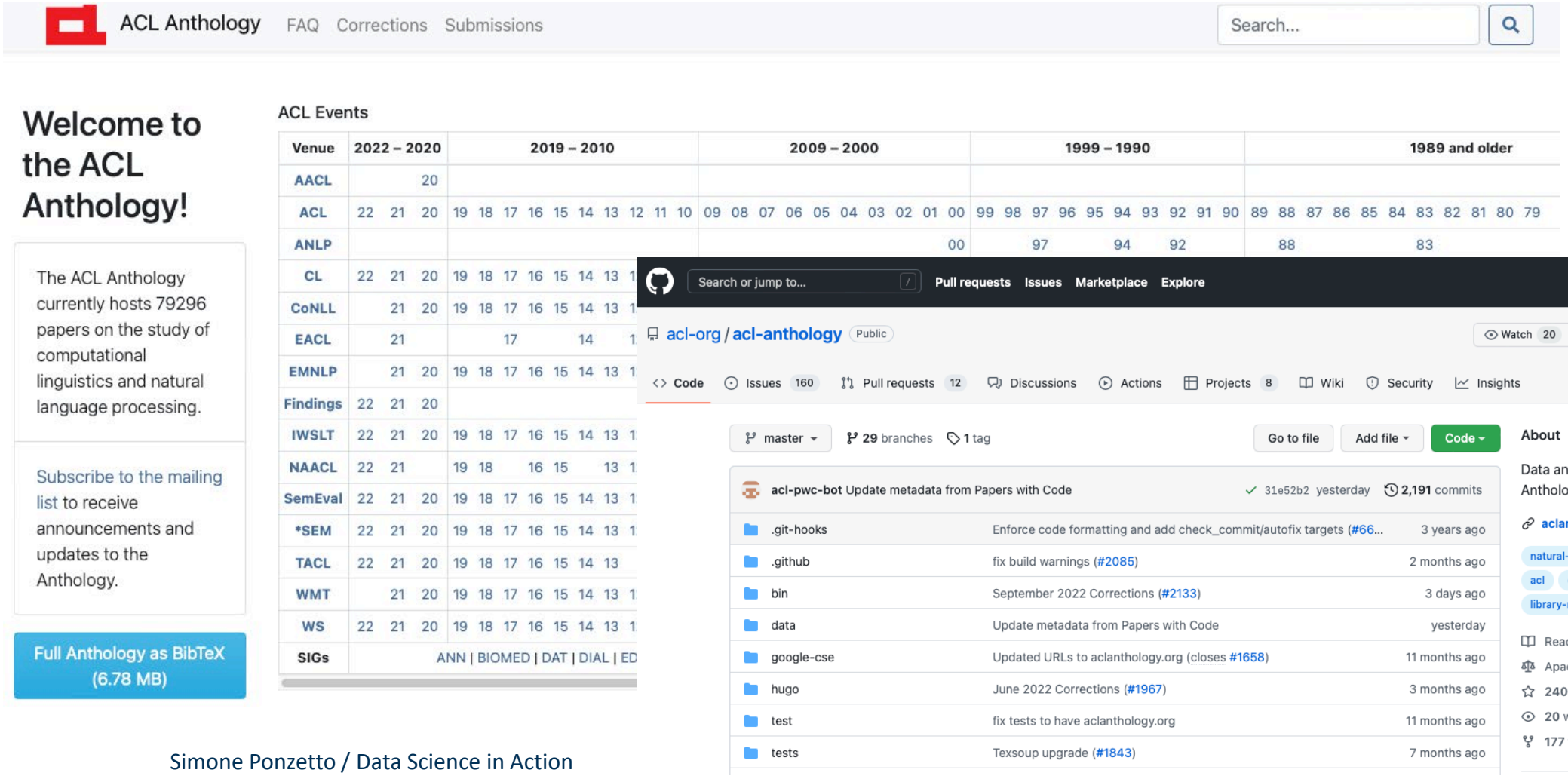
No releases published

**Packages**

No packages published

# ACL anthology

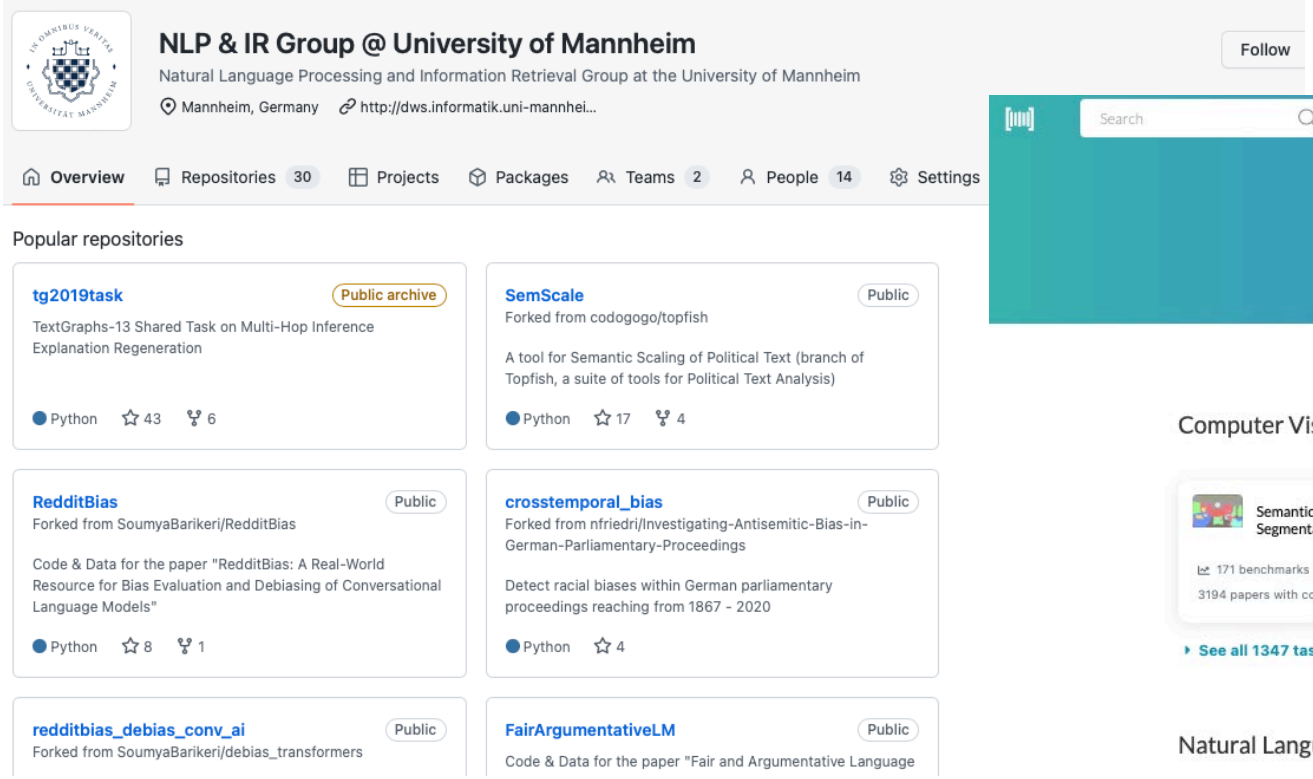
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



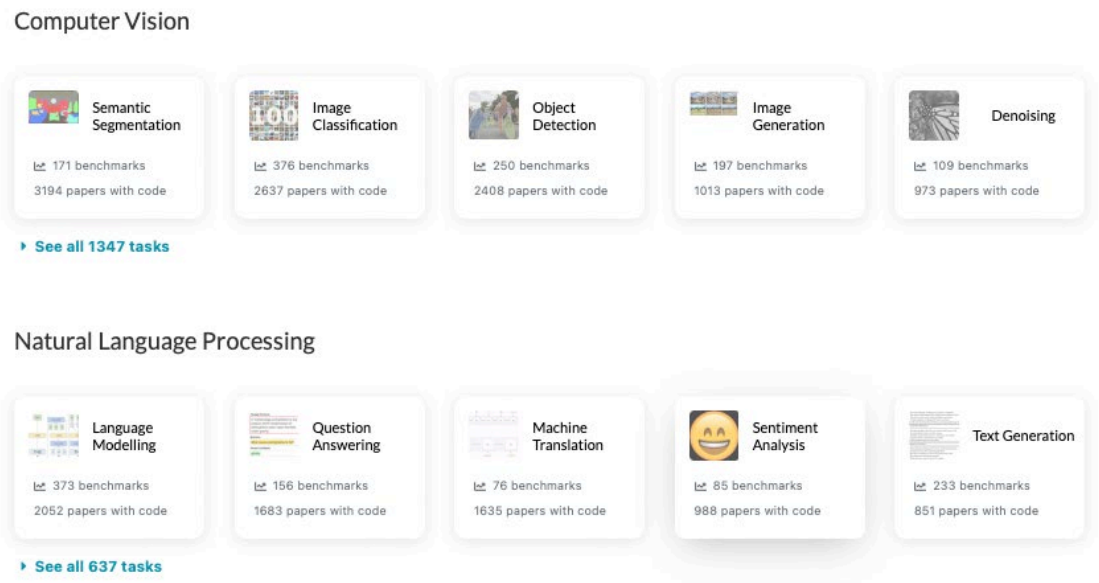

https://aclanthology.org

# Make the code available!

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

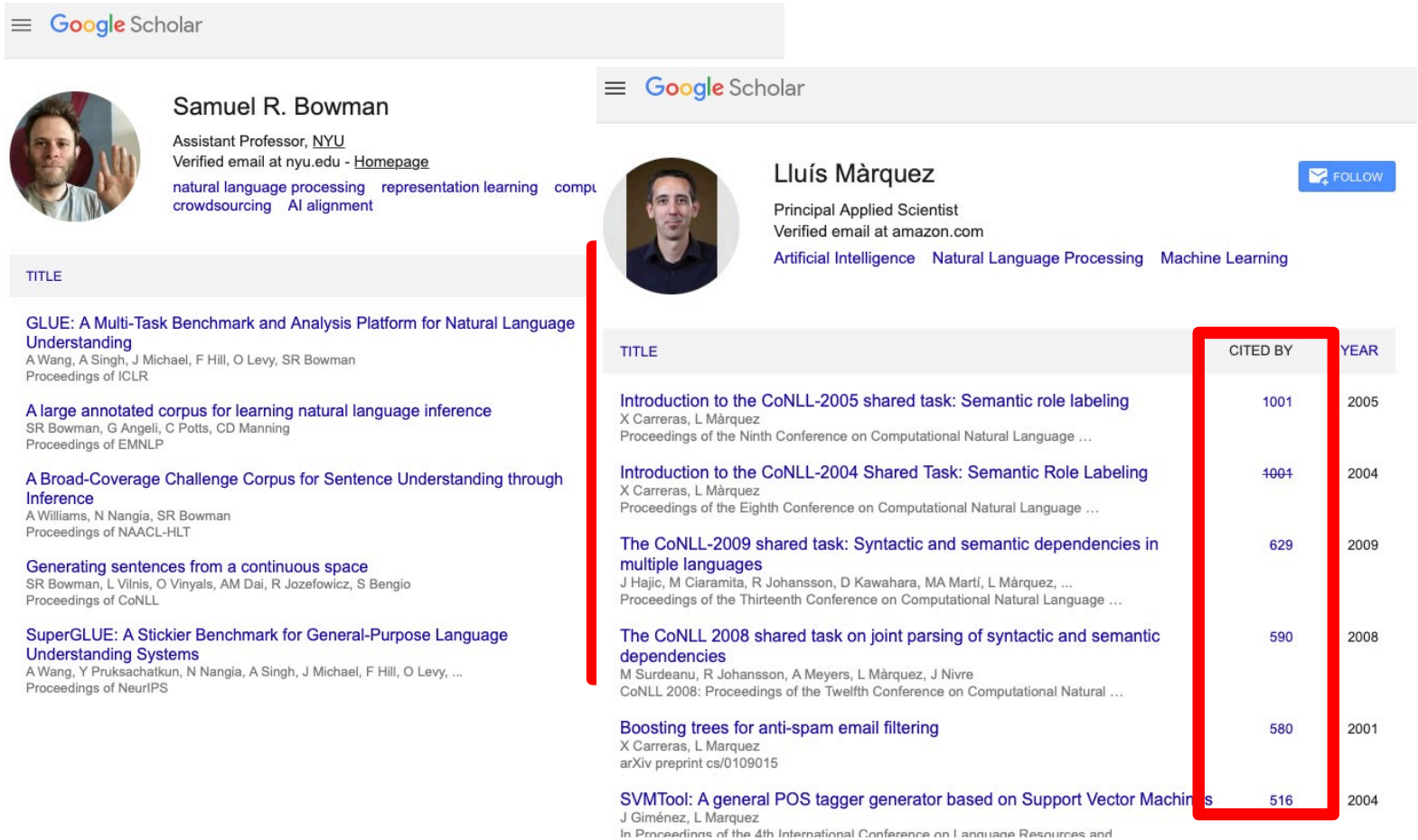







https://paperswithcode.com

# How to promote this: make it appealing for researchers




...

# Thanks! / Take-home messages

- Achieve **community-wide advancements** through the establishments of **openly shared experimental settings** (data, metrics) through *competition*
- ”Cement” the results into **openly available publications and code**
- Make it appealing through **visibility rewards** (most prominently, citations)
- Need to be aware of **limitations**! Example: “*rat-race chasing numbers*”