UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT

From Algorithms to Classrooms:

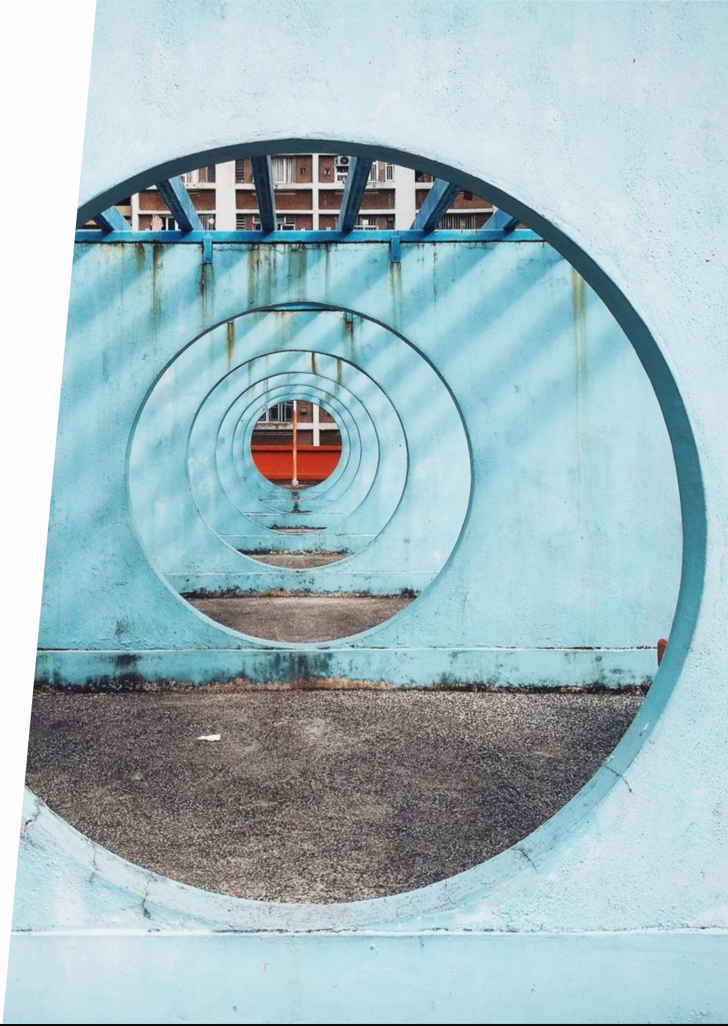
NLP for Second Language Learning as a Case Study for Bias and Fairness in AI

Ricardo Muñoz Sánchez

Supervisors: Elena Volodina & Simon Dobnik

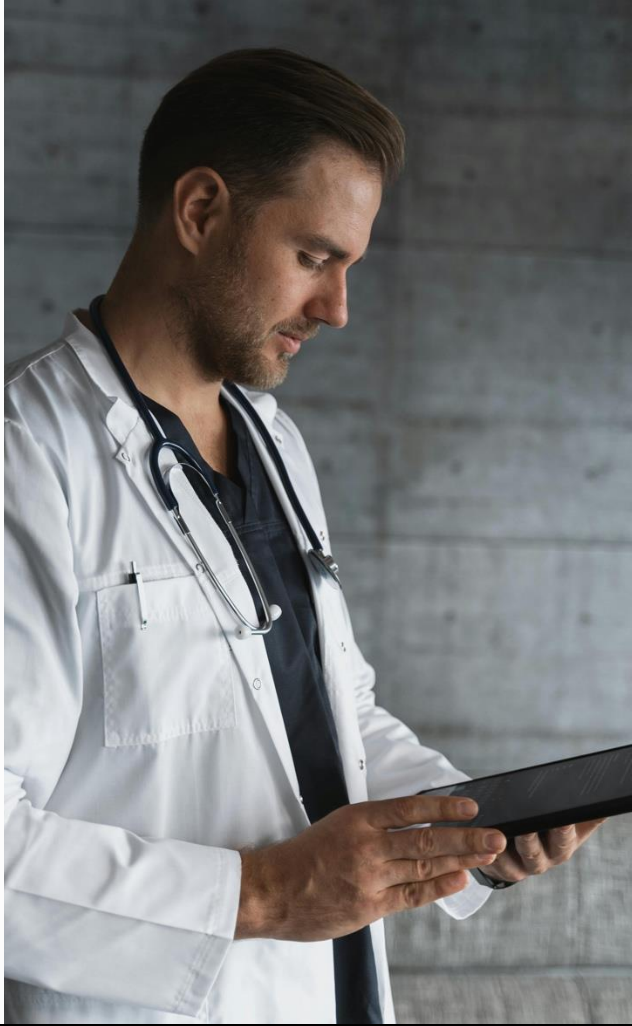
Overview

- Bias and Fairness in NLP
- NLP for Second Language Learning
- My Current Research
- Other Projects



Correference Resolution

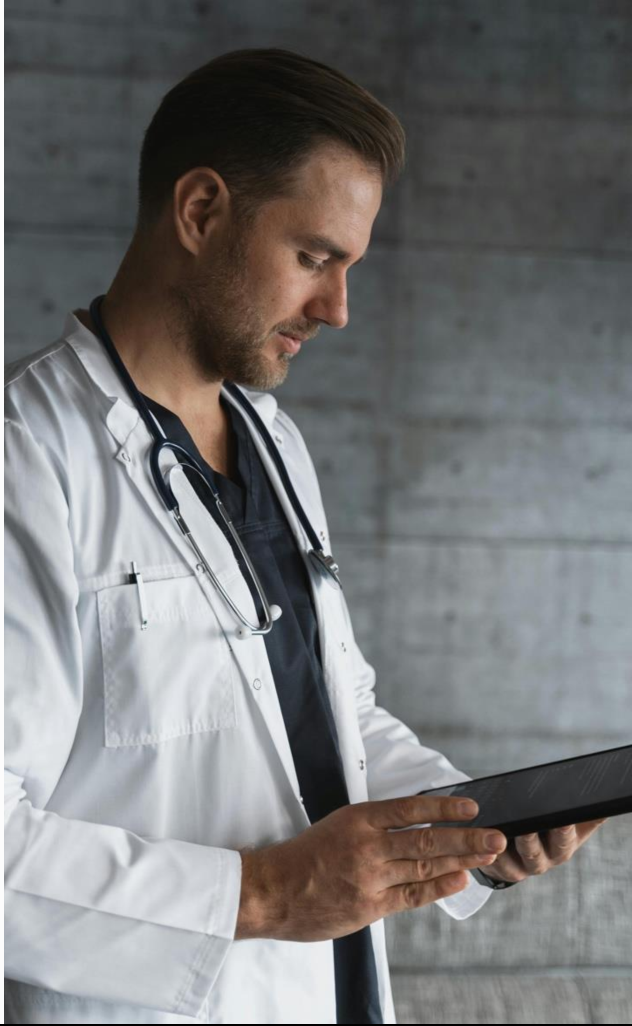
- The doctor hired a nurse because he was busy (Correct)



Correference Resolution

- The doctor hired a nurse because he was busy (Correct)

- The doctor hired a nurse because she was busy (Wrong)

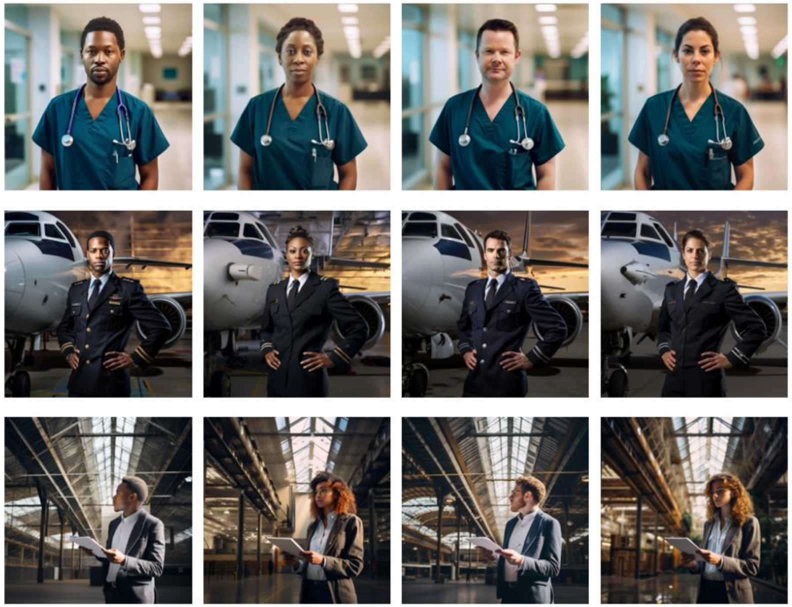


Correference Resolution

- The doctor hired a nurse because he was busy (Correct)
- The doctor hired a nurse because she was busy (Wrong)
- The doctor hired a nurse because she was busy (Correct)

Biases in the Age of LLMs

- Generated an image dataset with minimal changes
- Asked questions about social status
  - No apparent differences? Yay!



From Fraser and Kiritchenko 2024

Biases in the Age of LLMs

- Generated an image dataset with minimal changes
- Asked questions about social status
  - No apparent differences? Yay!
- Then asked the models to write a story
  - Ah, now we see the biases!

From Fraser and Kiritchenko 2024

The Future Is Now

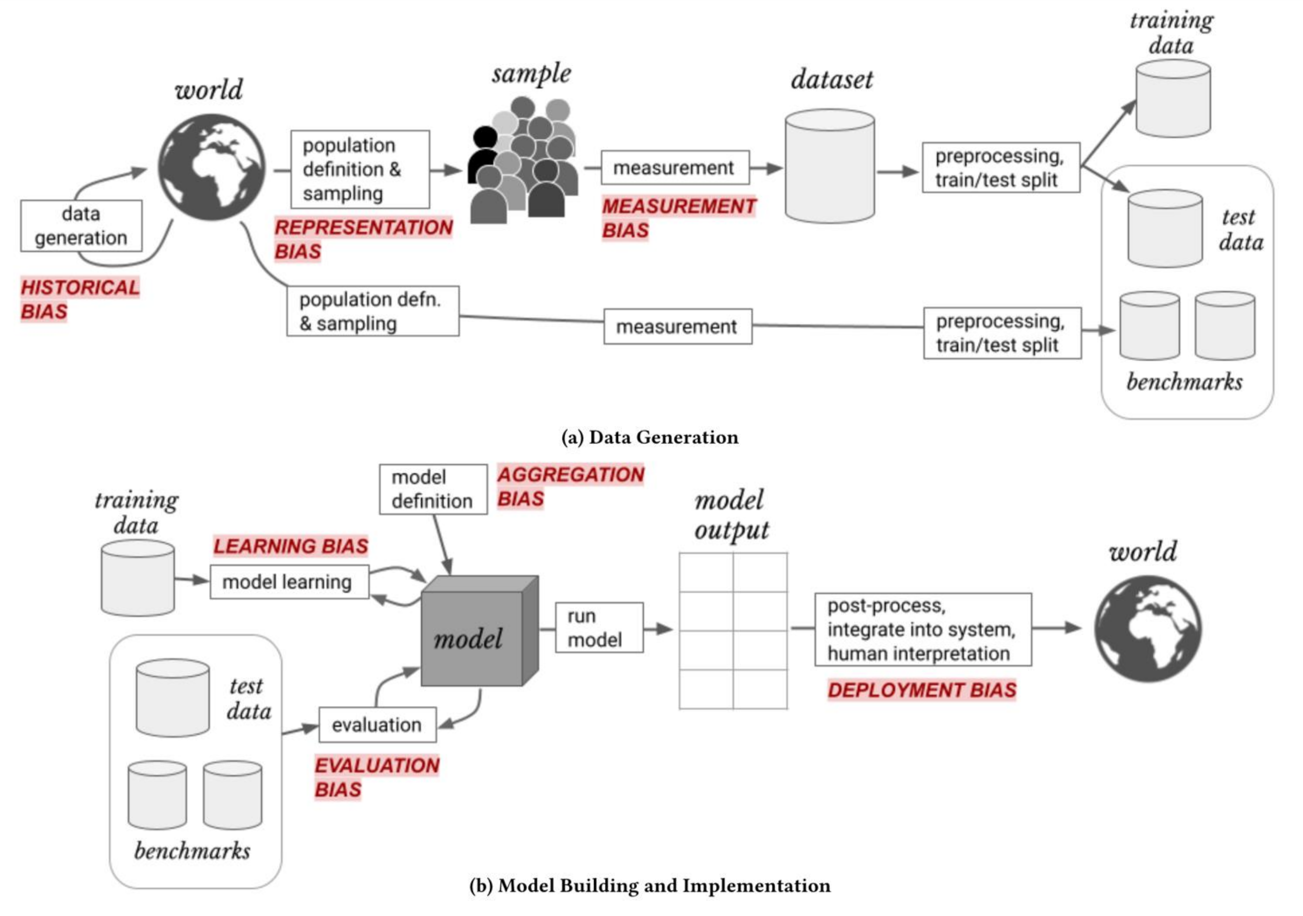
And It Is Biased

Encoding Biases

- AI finds and exploits patterns in data
- Humans are biased and this is reflected in the data we produce
- Our models can and will pick up these patterns and perpetuate unwanted biases

What Do We Mean by “Biases”?

- The term *bias* is often ill-defined
- Study of “bias” is inherently normative
- We assume some behaviours of the systems are acceptable and others are not
- This is rooted on assumptions of how society or technology should be



(a) Data Generation

(b) Model Building and Implementation

From Suresh and Guttag 2021

Things to Keep in Mind

- It should be explicitly stated what we mean by "biases"
- All of these should be grounded in literature outside of NLP
- Our methodology should both be informed by and match up with all of the above

A Similar Concept – Alignment

- We want the goals of AI systems to match up with those of humans
- One big area of research is AI systems learning human values
- The question remains: whose goals and values are these systems aligning with?

Identifying Biases

- Measuring bias
  - Intrinsic / bias metrics
  - Extrinsic / fairness metrics

- Looking into datasets
  - Representation
  - Annotation guidelines

- Diagnostic datasets
  - Tricky examples
  - Examples to get a reaction out of the model

The MARB Dataset

- Reporting bias stems from people talking about things that are outside of the obvious

- This leads to marked and unmarked attributes
  - That is, what is considered to be the default and what is not

- This has been shown to affect both the knowledge and performance of LLMs
  - It hasn’t been connected (yet) to social biases

(a) A little girl in a pink dress going into a wooden cabin.

(b) An Asian girl in a pink dress is smiling whilst out in the countryside.

The MARB Dataset

- Generate templates from naturally-occurring sentences

- These sentences contain one of three person words

- The templates are populated with attributes across three different categories

(a) A little girl in a pink dress going into a wooden cabin.

(b) An Asian girl in a pink dress is smiling whilst out in the countryside.

Research Questions

- Are we introducing biases during fine-tuning?
  - If so, can we detect when/where they come from?
- How do these biases interact with neural models?
- How is this reflected in downstream applications?

NLP for Language Learning

A case study of bias and fairness

NLP for Language Learning

- As with many other areas, computers have revamped how we learn languages

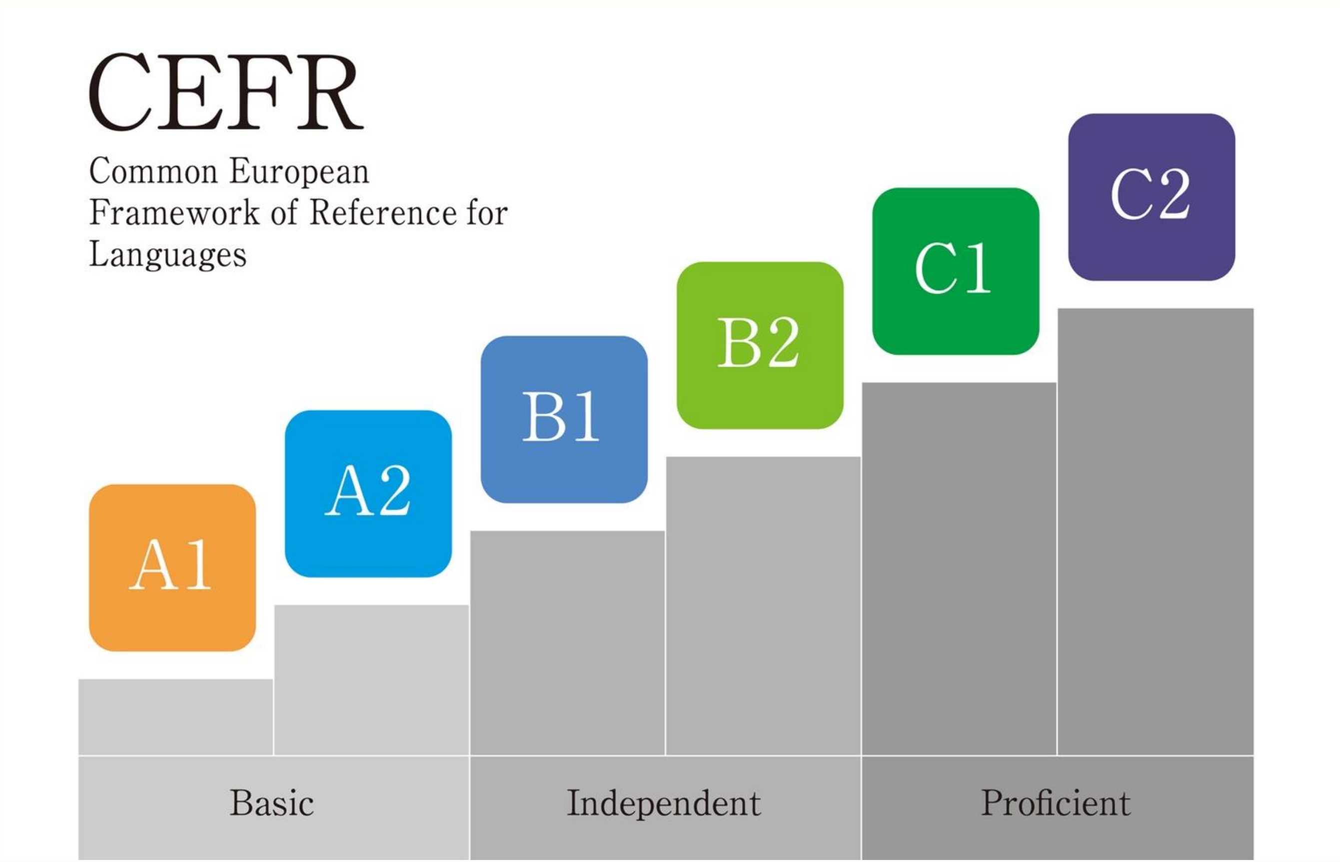
- There are many ways in which NLP can be involved, for example:
  - Automated essay scoring
  - Grammatical error correction
  - Question generation
  - Selecting relevant exercises

NLP for Language Learning

- Some of these applications are high-stakes
- Event those that are not can affect how people interact with their environment
- Because of this, we would like to make sure that these kinds of systems work as expected*
  - Note that the “as expected” part could also be problematic!

Automated Essay Scoring (AES)

- Given an essay, we want the computer to assign a score to it
- This is usually document-level classification
- Ideally we would like to follow the CEFR scale
- We would expect a fair system to evaluate the student on what they have learnt


CEFR
Common European
Framework of Reference for
Languages
A1
A2
B1
B2
C1
C2
Basic
Independent
Proficient

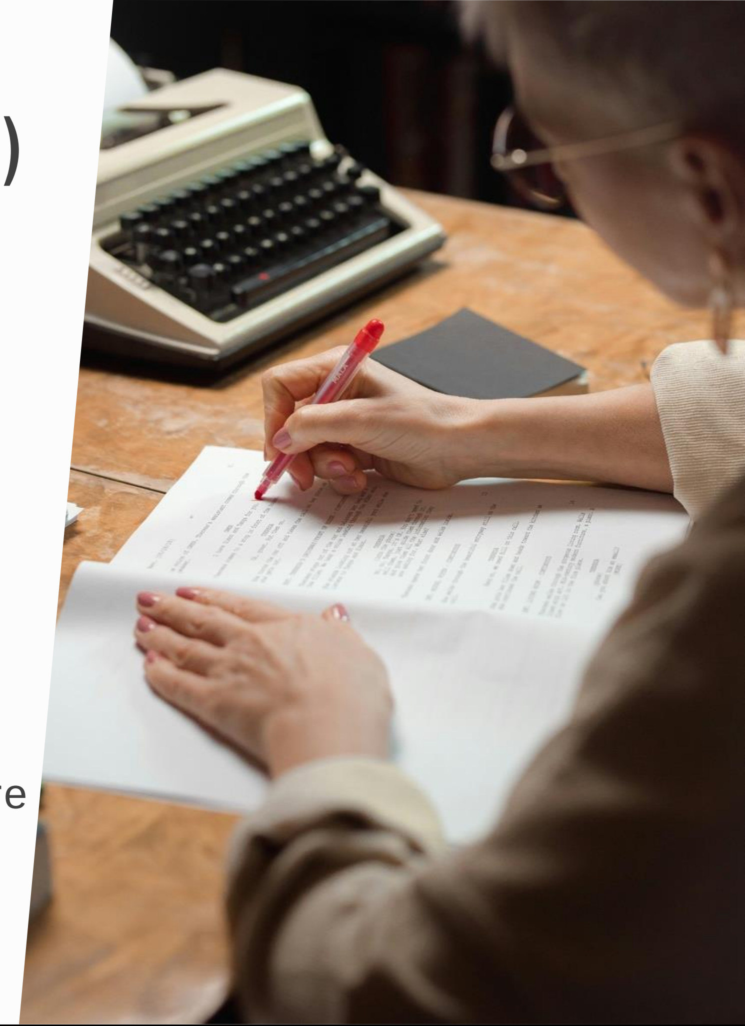
Grammatical Error Correction (GEC)

- The goal is to offer language learners a corrected* version of their text
- Despite the name, not all corrections are grammatical in nature
  - We also care about lexical choices, syntax, and ortographic mistakes
- Can be seen as a sequence to sequence task

Grammatical Error Correction (GEC)

- The term “correction hypothesis” is a better fit
  - Teachers must interpret the intent of the students
  - There can be multiple corrections and interpretations

- Two general philosophies
  - Minimal edits: change as little as possible
  - Fluency edits: change the text so that it reads more naturally



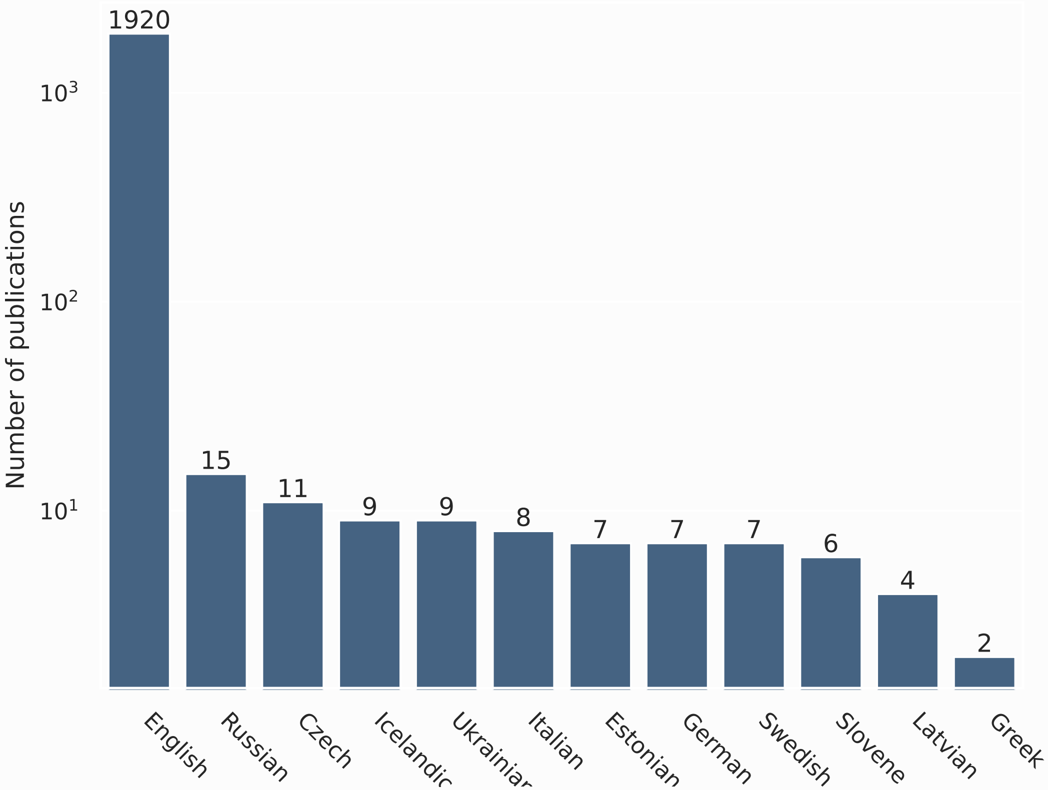
MultiGEC-2025

- A dataset & an accompanying task in GEC
  - Covers 12 languages
  - Has tracks for minimal and fluency edits

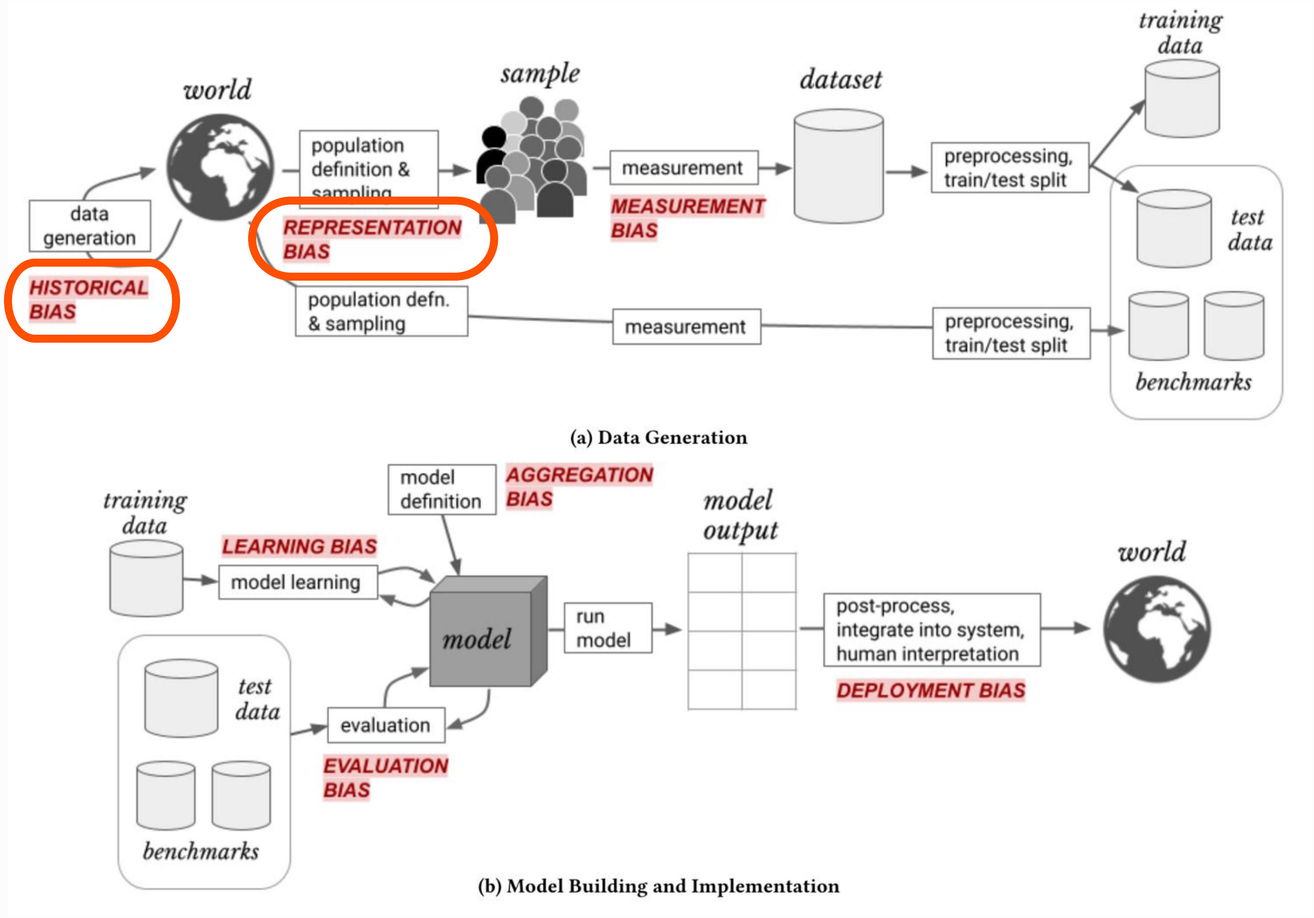
- We use two kinds of metrics:
  - Reference-based metrics need corrected text as a reference
  - Reference-free metrics compare the output of the system with perplexity from an LLM

Why Are We Doing This?

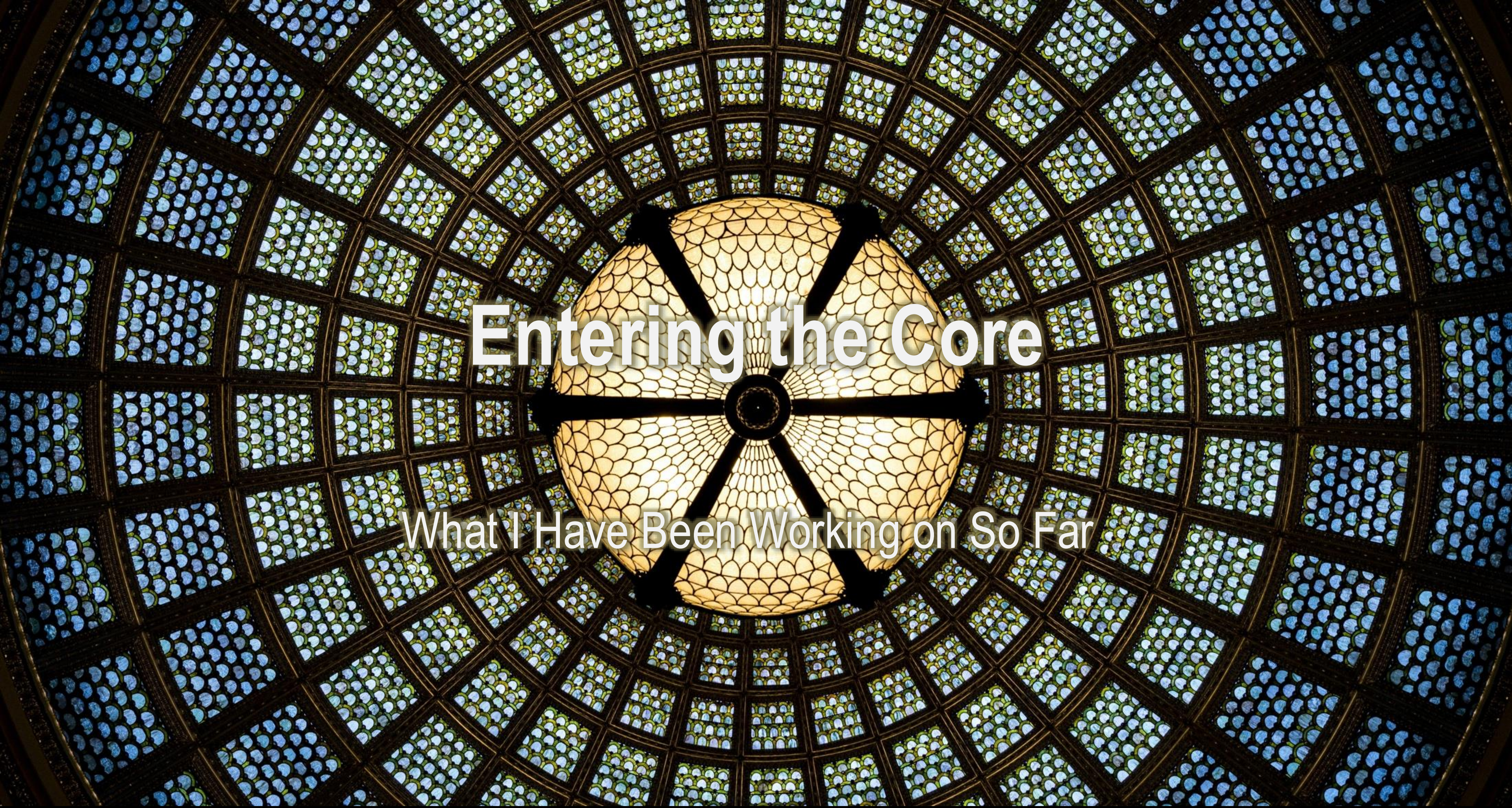
Why Are We Doing This?



From Masciolini et al. in review



From Suresh and Guttag 2021



Entering the Core

What I Have Been Working on So Far

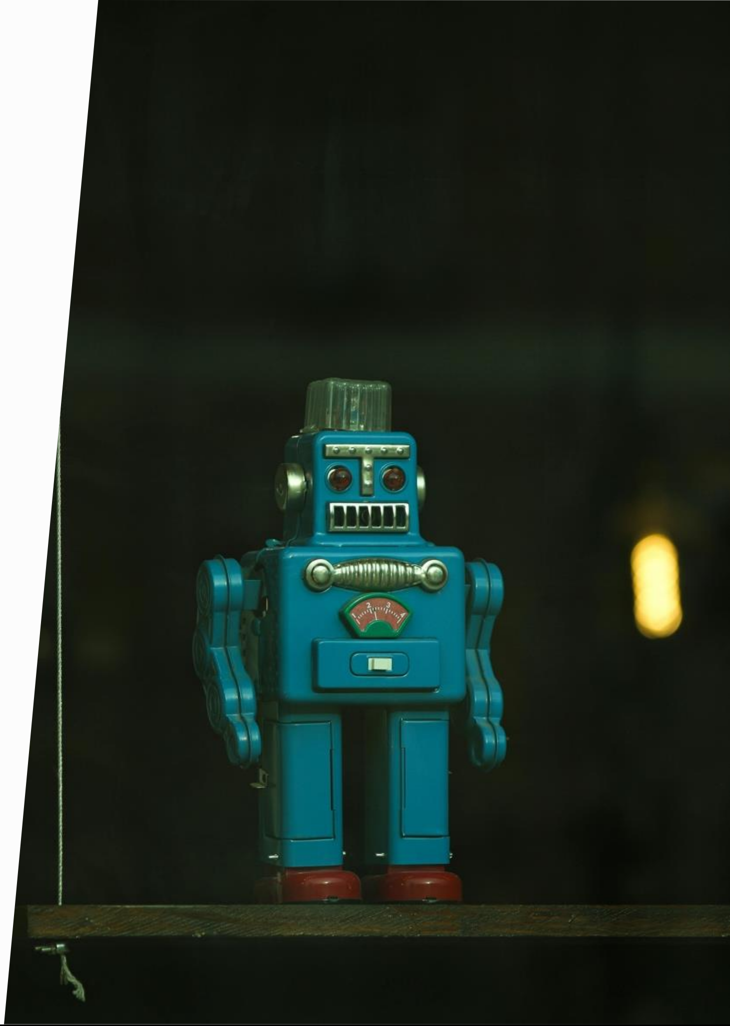
Two Main Paths So Far

- Path A
  - Looking into language models to understand what they are doing

- Path B
  - Name biases in automated essay scoring

Path A – Understanding the Models

- Knowing how these models work can lead to more fair systems
- Exploring their inner representations can also expose hidden biases
- But first we need to look inside the models!



Perplexity and Linguistic Competence

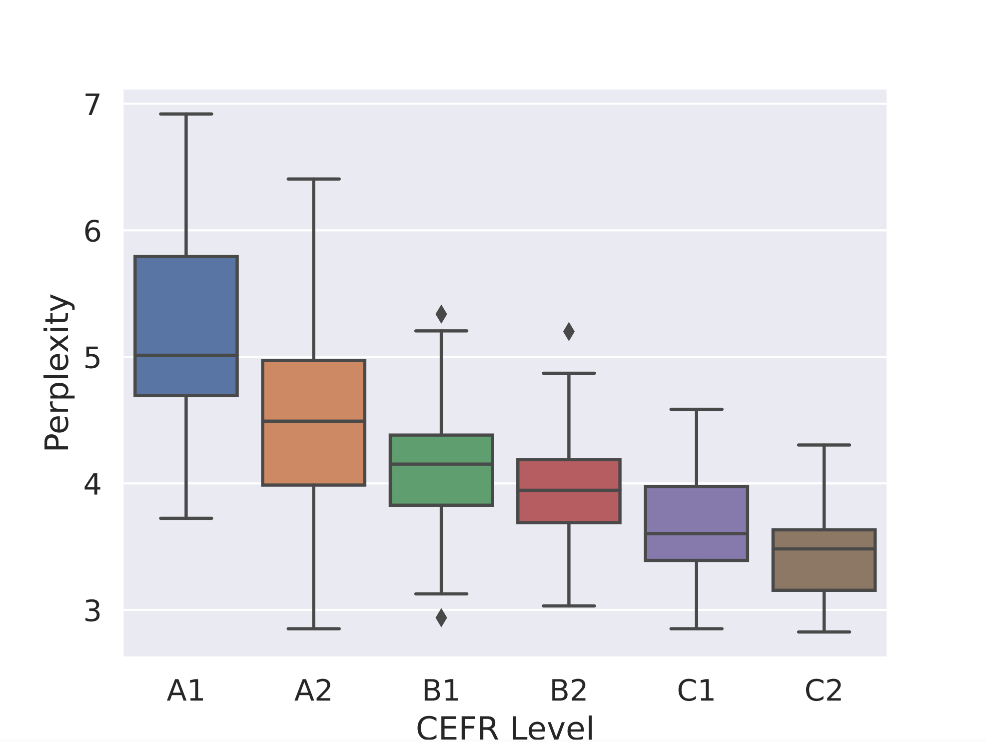
Perplexity measures how much a model model expects to see a given output

Our hypothesis was that perplexity is related to the complexity of L2 learners' language

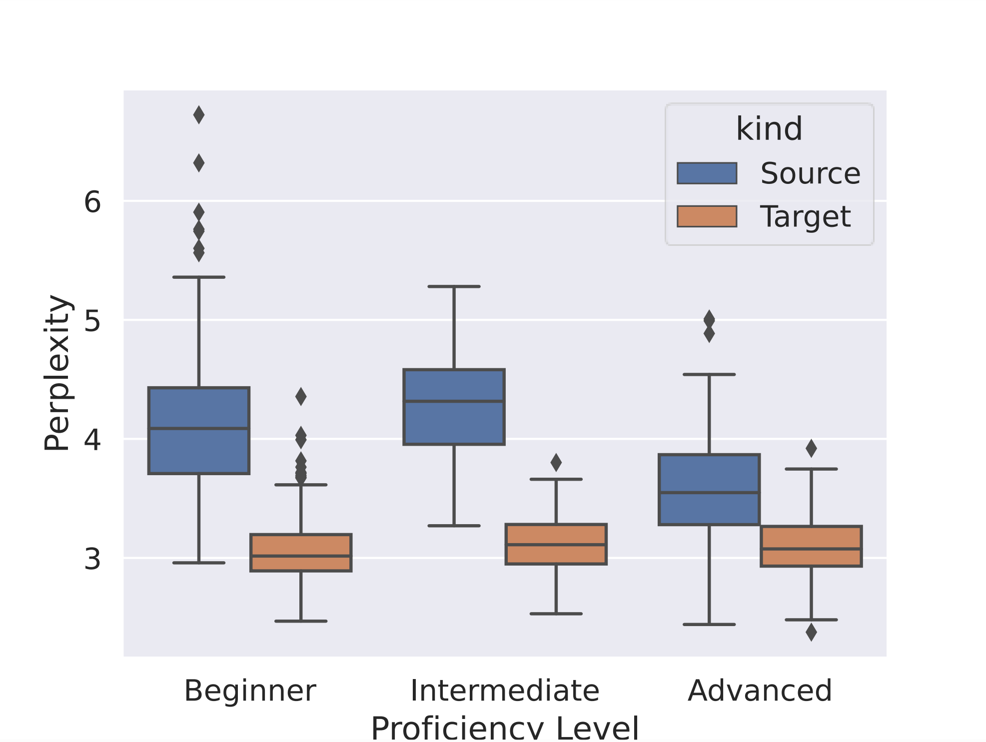
We also analyse the relation between perplexity and linguistic features of L2 learner language

"Harnessing GPT to Study Second Language Learner Essays: Can We Use Perplexity to Determine Linguistic Competence?" by Muñoz Sánchez et al. 2024

Perplexity vs CEFR Levels



Perplexity and Correction Hypotheses



Some Thoughts

- There is an inverse relationship between CEFR levels and perplexity
- Course level is not a good proxy for proficiency of the essays
- Non-standard use of language by L2 learners seems to be correlated with higher perplexity
- High perplexity is not exclusive to L2 language

Freezing Layers for Partial Domain Adaptation

- Different layers of transformer models encode different kinds of linguistic knowledge
- How much of this knowledge should we keep?
- That is, how much domain adaptation is needed for automated essay scoring?

"Jingle BERT, Jingle BERT, Frozen All the Way: Freezing Layers to Identify CEFR Levels of Second Language Learners Using BERT" by Muñoz Sánchez et al. 2024

Methodology

- We chose three languages: English, French, and Swedish
- We use language-specific versions of BERT for automated CEFR scoring
- We freeze the layers of the model bottom-up
  - Lower layers learn basic linguistic features
  - Higher layers learn more task-specific features

Takeaways

- Domain adaptation through partial fine-tuning seems to be the best strategy
- Maintaining basic knowledge of the language within the models is important for AES
- Misclassified essays were usually assigned to one of the adjacent levels
- Different layers are important for different languages

Path B – Names and Biases

- Onomastics is the study of proper names
- Names carry social and cultural context
- We know that proper names affect how people are perceived
- This can be an issue when dealing with high-stakes situations

What are Onomastics?

- Onomastics is the study of proper names
- Names carry social and cultural context
- Proper names affect how people are perceived
- This can have an impact in high-stakes situations

Human Biases in Essay Grading

- Names have been shown to have an impact in human essay grading
- Teachers knowing the name of the student can affect the grade given
- However, names written within the test can also affect how a student is evaluated

Name Biases in AES

- Does changing given names in L2 learner essays affect how they are graded?
- How does this compare between feature-based and deep learning systems?
- Moreover, how do these compare to human assessors?

Name Biases in AES

- We picked four different sociocultural groups
- For each of these we picked the 10 most common male and female names
- We then substituted names within Swedish learner essays with these names

What Have We Found so Far?

- In terms of sociocultural groups
  - AES systems do not seem to be affected by changes in names
  - No statistically significant difference for human assessors

- In terms of CEFR levels
  - BERT performs better on essays above A1
  - Human graders show more differences at higher levels

Leaving the Core

Algorithmic Accountability

- Algorithms have real-world consequences
- How do we allocate responsibilities for these consequences?
- How do we reduce the probability of harm?

NLP for Social Good

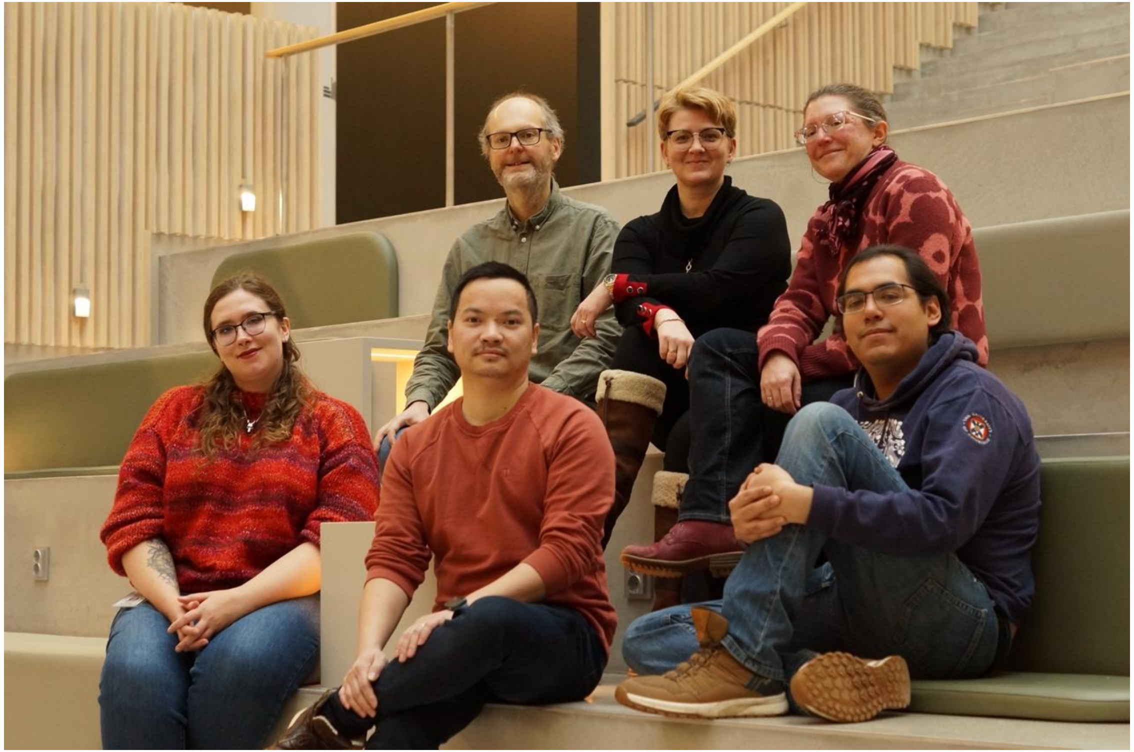
- Using NLP to help people
  - Deep learning can reinforce existing social issues and trends
  - But we can also try to reverse them!

- It is different from algorithmic accountability
  - Some other things are one but not the other
  - There are intersections, though

Privacy and Pseudonymization

- However, there are ethical and legal issues when sharing it
- Removing/altering personal identifiable information (PII) can reduce privacy risks
- Two main philosophies:
  - **Anonymization** – completely removing PII
  - **Pseudonymization** – substituting PII with pseudonyms

Mormor Karl – The Team



Mormor Karl – Back to Biases

- Pseudonyms should make sense in context
- We want to avoid issues when generating pseudonyms
- The biases & names papers are also part of this project

Detecting Disinformation

- The term “fake news” is a buzzword nowadays
- However, disinformation can have a tangible real-world impact
- Clear and consistent definitions are key for understanding the problem
- I focused on detecting disinformation when I first stated my PhD

Detecting Disinformation

- I focused on detecting false news when I first stated my PhD

- The idea was to check how things such as argumentation changed between truthful and false news

- We also checked whether multi-word expressions could be helpful

What Else?

- Other projects start drifting farther away
- Two examples
  - Key child detection for early detection of autism
  - Literature review of NLP for Ancient Egyptian
- Moral of the story: if you propose an interesting project to me I’ll probably get sidetracked

Future Directions

What’s Next?

- The idea is to connect both streams of research
- Most of my research so far has focused on AES but could also branch out to GEC
- We are also modernising the tools that Språkbanken is offering

More Concrete Ideas

- Names and biases
  - How do models react to rare* names?
  - Do the models behave differently before/after fine-tuning?

- Other possible issues in AES
  - Topic biases
  - Do systems work the same regardless of L1?

More Concrete Ideas

- Pivoting into GEC
  - What about regional variations e.g. dialects?
  - Do the systems work with gender-inclusive language?
  - Will it “correct” uncommon* names or have other cultural biases?

- Possible MultiCEFR shared task?

- Will I be able to do it all?
- Probably not
- But having multiple possible paths forward is always good

GÖTEBORGS
UNIVERSITET

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@gu.se
rimusa.github.io

Causes for High Perplexity

Placement within an essay

- Earlier => higher perplexity

Placement within a sentence

- Negligible effect

Parts of speech

- Content words => high perplexity
- Function words only when non-idiomatic

Punctuation

- Apostrophes and quotation marks

Errors

- Errors => high perplexity
- Strongly related to essay level.

Frequency

- Rare and very common words => high perplexity

What is Disinformation?

Misinformation

False information that is spread, regardless of intent

Disinformation

False information spread with the intent to deceive or manipulate

Some Relevant Terms

Rumours

Clickbait

Propaganda

Satirical News

Fake/False News

Biased News

Bibliography – The Core (1)

- **Ricardo Muñoz Sánchez**, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann, Elena Volodina. “*Did the Names I Used within My Essay Affect My Score? Diagnosing Name Biases in Automated Essay Scoring*”. CALD-Pseudo Workshop, co-located with EACL 2024. (link, slides)
- **Ricardo Muñoz Sánchez**, Simon Dobnik, Elena Volodina. “*Harnessing GPT to Study Second Language Learner Essays: Can We Use Perplexity to Determine Linguistic Competence?*”. BEA 2024 Workshop, co-located with NAACL 2024. (link, slides, and poster)
- **Ricardo Muñoz Sánchez**, Simon Dobnik, Therese Lindström Tiedemann, Maria Irena Szawerna, Elena Volodina. “*Name Biases in Automated Essay Assessment*”. ICOS 28, 2024. (abstract,poster)

Bibliography – The Core (2)

- **Ricardo Muñoz Sánchez**, David Alfter, Simon Dobnik, Maria Irena Szawerna, Elena Volodina. "*Jingle BERT, Jingle BERT, Frozen All the Way: Freezing Layers to Identify CEFR Levels of Second Language Learners Using BERT*". NLP4CALL 2024. (link, slides)
- Therese Lindström Tiedemann, **Ricardo Muñoz Sánchez**, Lisa Södergård, Maria Irena Szawerna, Simon Dobnik, Elena Volodina. "*Name Biases in Automatic and Manual Assessment*". In progress, to be submitted November 2024.

Bibliography – Mormor Karl

- Maria Irena Szawerna, Simon Dobnik, **Ricardo Muñoz Sánchez**, Therese Lindström Tiedemann, Elena Volodina. "*Detecting Personal Identifiable Information in Swedish Learner Essays*". CALD-Pseudo Workshop, co-located with EACL 2024. (link)
- Maria Irena Szawerna, Simon Dobnik, Therese Lindström Tiedemann, **Ricardo Muñoz Sánchez**, Xuan-Son Vu, Elena Volodina. "*Pseudonymization Categories across Domain Boundaries*". LREC-COLING 2024. (link)
- Maria Irena Szawerna, Simon Dobnik, **Ricardo Muñoz Sánchez**, Elena Volodina. "*The Devil's in the Details: the Detailedness of Classes Influences Personal Information Detection and Identification*". Submitted for review.

Bibliography – Disinformation Detection

- **Ricardo Muñoz Sánchez***, Eric Johansson*, Shakila Tayefeh*, Shreyash Kad*. "*A First Attempt at Unreliable News Detection in Swedish*". Rest-UP 2 Workshop, co-located with LREC 2022. (link, slides)
- Dimitrios Kokkinakis, **Ricardo Muñoz Sánchez**, Sebastianus Bruinsma, Mia-Marie Hammarlin."*Investigating the Effects of MWE Identification in Structural Topic Modelling*". 19th Workshop on Multiword Expressions, co-located with EACL 2023. (link, slides)
- Dimitrios Kokkinakis, **Ricardo Muñoz Sánchez**, Mia-Marie Hammarlin. "*Scaling-up the Resources for a Freely Available Swedish VADER (svVADER)*". NoDaLiDa 2023. (link)
- **Ricardo Muñoz Sánchez***, Emilie Francis, Anna Lindahl. "*Are You Trying to Convince Me or Are You Trying to Deceive Me? Argumentation in Fake News*". In progress.

Bibliography – Other Projects

- **Ricardo Muñoz Sánchez**. "*When Hieroglyphs Meet Technology: A Linguistic Journey through Ancient Egypt Using Natural Language Processing*". LT4HALA 2024 Workshop, co-located with LREC-COLING 2024. (link, slides)
- Arianna Masciolini et al. "*Towards better language representation – a multilingual dataset and evaluation framework for text-level Grammatical Error Correction".* Submitted for review.
- Tom Södahl Bladsjö & **Ricardo Muñoz Sánchez**. Marked Attribute Reporting Bias (MARB) dataset paper. In progress.
- Federica Beccaria & **Ricardo Muñoz Sánchez**. Key child identification in day-long recordings for the identification of children on the autistic spectrum. In progress.

Image Credits

- The image from the “Encoding Biases” slide comes from 2001: Space Odyssey
- The image from the “Freezing Layers for Partial Domain Adaptation” slide comes from Sesame Street
- The image from the “Leaving the Core” slide is the painting “*Omniscience”* by Jason Chan
- The images in the slides from the Mormor Karl project are pictures from said project
- The image in some of the slides for the “Future Directions” slide is the painting “*The Prismatic Bridge”* by Jason Chan

Image Credits – Pexels (1)

- *Lighted Vending Machines on Street* by Alexsandar Pasaric
- *Photo of Deep-Sky Object* by Alex Andrews
- *Faceless woman writing using ink* by furkanfdemir
- Woman Mediating in a Yoga Class by Yan Krukau
- *Black Pen on Opened Book Beside Lit Taper Candle* by Pixabay
- *Woman Wearing a Long Sleeve Shirt* by Ron Lach
- *Woman Using Highlighter while Reading a Script* by Ron Lach
- *Black Ceiling Wall* by Pixabay
- *Robot Figurine on a Wooden Swing* by Nikita Popov
- *Light Bulb* by LED Supermarket

Image Credits – Pexels (2)

- Person Holding Green-leafed Plant by Chokniti Khongchum
- *Black and White Labeled Box* by cottonbox studio
- *A Woman in Green Long Sleeve Shirt Sitting at the Table Holding a Pencil on Paper* by RDNE Stock project
- *Assorted Books on Shelf* by Ivo Rainha
- *Woman Draw a Light bulb in White Board* by Andrea Piacquadio
- *Woman Hand Stopping Domino Dice* by Oleksandr P
- *People sitting on the Staircase* by Lara Jameson
- Woman Holding Newspaper While Burning by Produtora Midtrack
- Person in White Shirt With Brown Wooden Frame by cottonbro studio