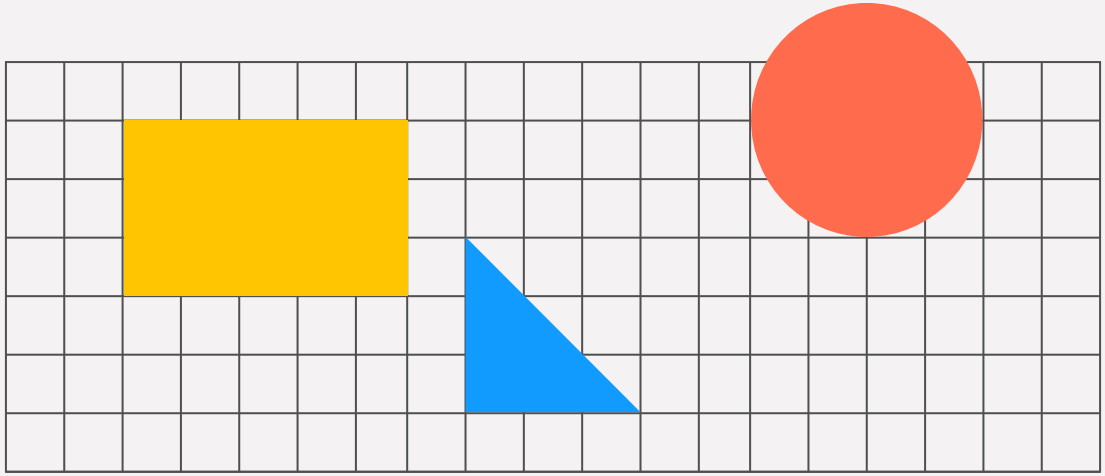
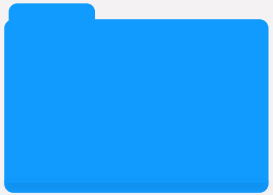
ENLP Projects

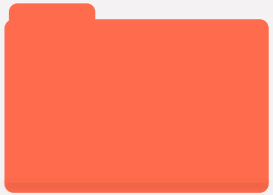
2025 Spring
Professor Nathan Schneider
TA: Xiulin Yang & Blake Wang



BabyLM



Morphological Inflection



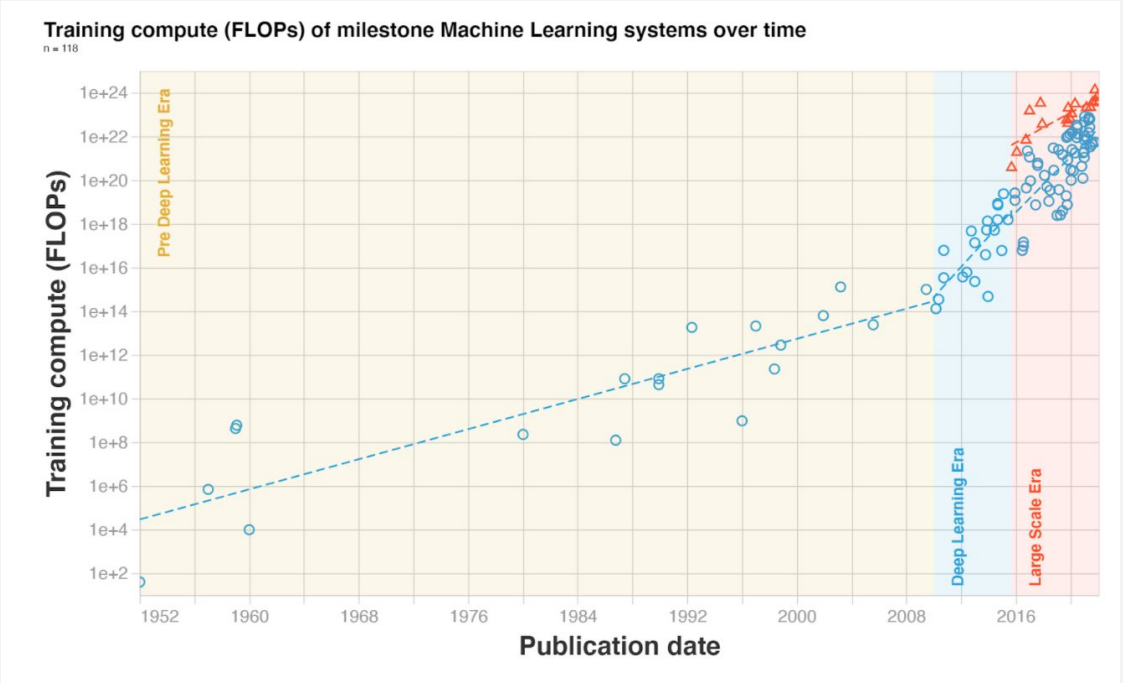
Lexical Simplification

Legal NER

https://babylm.github.io/archive_2023.html

BabyLM Challenge

Background & Motivation: For language models to achieve impressive results, they must be trained on hundreds of times more linguistic input than a typical human encounters in a lifetime!



BabyLM Challenge

- **Task:** In this shared task, participants are challenged to train a language model from scratch with the same limited linguistic input that a child receives.
- **Data**: 10M words (strict-small track) or 100M words (strict track)
- **Evaluation**: the evaluation pipeline is provided in the github (https://github.com/babylm/evaluation-pipeline-2023)
    - Linguistic knowledge evaluation (BLiMP) https://github.com/nyu-mll/jiant/tree/blimp-and-npi
    - Semantic understanding (Super)GLUE

Ideas for research directions

- What pretraining paradigms helps LMs to learn and generalize?
  - Curriculum learning (Oba et al., 2023)
  - Introducing inductive bias (Papadimitriou & Jurafsky, 2023)
- Does adding multimodal data help? (Klerings et al., 2024, Kuribayashi & Baldwin, 2025)
- Does perplexity align with LMs' targeted evaluation task? (Xu et al, 2025)
- …

SIGMORPHON–UniMorph Shared Task on Typologically Diverse and Acquisition-Inspired Morphological Inflection Generation

Background & Motivation: Existing language models perform impressively in high-resource languages but still lag behind in low-resource settings. Developing a high-quality inflection model can also contribute to building the UniMorph dataset.

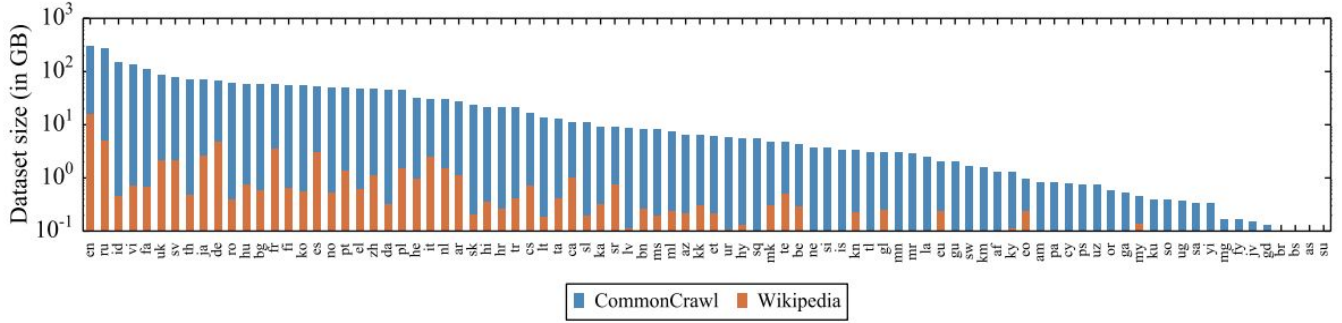


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

https://github.com/sigmorphon/2023InflectionST

SIGMORPHON–UniMorph Shared Task on Typologically Diverse and Acquisition-Inspired Morphological Inflection Generation

- In this shared task, participants will develop a model that generates morphological inflections from a given lemma and a set of morphosyntactic features specifying the target form.
- This task is suitable for students with a linguistic background, so make sure your team includes at least one **linguistics** student!
- Data is provided in the repository.
- Evaluation: P, R, F, and Accuracy

Spanish	cancelar	V;IMP;ACC(2,SG);NOM(INFM,2,SG)	cancélate
Turkish	asimptot	N;NOM(PL;PSS(1,PL))	asimptotlarımız

Ideas for research directions

- How do multilingual morphological models compare to monolingual ones in terms of accuracy and generalization? Is it always the case the one is better than the other? Try to experiment with a few languages.
- What linguistic/typological features help the transfer learning? (e.g., Lin et al., 2019)
- How can we develop a robust multilingual inflection model that generalizes across typologically diverse languages? (e.g., Wu et al., 2021, Cotterell et al., 2018)
- The shared-task 2023 finding paper (Goldman et al., 2023)
- Some useful repositories: https://github.com/CUNY-CL/yoyodyne (a framework for small-vocabulary sequence-to-sequence generation); OpenNMT: https://github.com/OpenNMT/OpenNMT-py (an open-sourced neural machine translation)

Lexical Simplification

- **Background & Motivation**: Lexical simplification is the process to replace difficult words to the ones that are easier to read and understand. This task helps create texts that are friendly to foreign language learners, people with lower literacy levels or reading impairments.

- **Task**: Given a sentence and a complex word contained in it, return an ordered list of "simpler" valid substitutes.

That prompted the military to **deploy** its largest warship, the BRP Gregorio del Pilar, which was recently acquired from the United States.

output

Possible substitutes: send, use, move, position…

Lexical Simplification

- **Data:** training data not provided, but you can explore lexical datasets such as BenchLS and NNSEval (more datasets here)

- **Evaluation:** evaluation script available with ~300 test sentences
  (Do not include gold annotations from the test set in your training!)

  - Precision, accuracy, recall, F1-score
  - Mean Average Precision@K (how many predictions are relevant, and how well are they ranked?)
  - Potential@K (how many gold items are found in top K?)
  - Accuracy@K@top1(are the top K predictions found at the top of the gold list?)

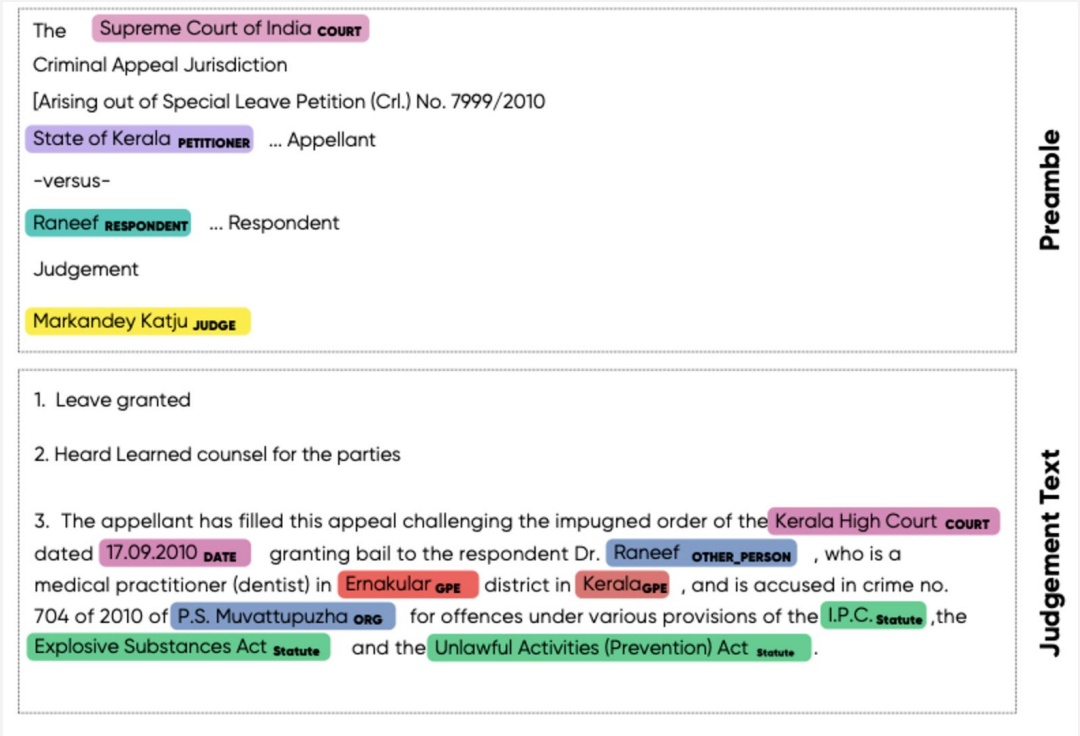
Ideas for research directions

- The evaluation metrics we use may not be perfect. What additional factors can be taken into consideration to capture the quality and usefulness of our model?
- Different users may have different needs when it comes to lexical simplification. How can we personalize the model to address the needs of different user groups? (North et al., 2022)
- This task focuses on the latter parts of the lexical simplification pipeline (e.g. substitute generation, selection, ranking, …). Can earlier components such as complex word identification (CWI) come into play? (Matthew et al., 2024)

https://github.com/Legal-NLP-EkStep/legal_NER

Legal NER

Background & Motivation: Legal documents have peculiar named entities like names of petitioner, respondent, court, statute, judge, etc. These entity types are not recognized by standard Named Entity Recognizer.



https://github.com/Legal-NLP-EkStep/legal_NER

Legal NER

Task: Extract legal named entity from legal documents (preamble, judgment).

Data: 9435 judgment sentences and 1560 preambles, sampled from 1950 to 2021

Evaluation: P, R, F, test set available

Named Entity	Extract From	Description
COURT	Preamble, Judgment	Name of the court which has delivered the current judgement if extracted from Preamble. Name of any court mentioned if extracted from judgment sentences.
PETITIONER	Preamble, Judgment	Name of the petitioners / appellants /revisionist from current case
RESPONDENT	Preamble, Judgment	Name of the respondents / defendents /opposition from current case
JUDGE	Premable, Judgment	Name of the judges from current case if extracted from preamble. Name of the judges of the current as well as previous cases if extracted from judgment sentences.
LAWYER	Preamble	Name of the lawyers from both the parties
DATE	Judgment	Any date mentioned in the judgment

...

Ideas for research directions

- What can be adapted from standard NER methods, and what can be done to incorporate domain-specific corpora?
- A set of predefined legal named entities categories is provided. How to utilize this? How can we assess per-category performance?
- It is important to capture document level context. Can we refer to other cases mentioned in the document during inference? (Kalamkar et al., 2022)

What we expect from you

- Propose a falsifiable research question with rigid experiment design and implementation
- Review relevant papers and provide theoretical support to your hypothesis and experiment design.
- Follow the required format and structure of an ACL paper.
- It's ok to have negative results! Negative results are important findings, too!
- You can use Huggingface models and experiment with the use of data
  - data manipulation strategies: data augmentation
  - integrating resources beyond the provided task data
  - changing the way the day is used, e.g. curriculum learning
- You can also try combining different model architectures
- Please do not simply compare off-the-shelf or fine-tuned/prompt-engineered models with the task dataset. We want you to test data/modeling ideas that are motivated by the problem.

Schedule

Date	Milestone
Tuesday Mar 25	**Submit project team with topic**
Tuesday Apr 10	**1-2 page proposal due**
Tuesday Apr 22	**Progress update, including lit review, due**
Tuesday Apr 29	**Project presentations**

References

- Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.
- Tatsuki Kuribayashi and Timothy Baldwin. 2025. Does Vision Accelerate Hierarchical Generalization in Neural Language Learners?. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1865–1879, Abu Dhabi, UAE. Association for Computational Linguistics.
- Alina Klerings, Christian Bartelt, and Aaron Mueller. 2024. Developmentally Plausible Multimodal Language Models Are Highly Modular. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 118–139, Miami, FL, USA. Association for Computational Linguistics.
- Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Oseki. 2023. BabyLM Challenge: Curriculum learning based on sentence complexity approximating language acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 290–297, Singapore. Association for Computational Linguistics.
- Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 Shared Task 0: Typologically Diverse Morphological Inflection. In Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 117–125, Toronto, Canada. Association for Computational Linguistics.
- Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the Transformer to Character-level Transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.
- Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 Shared Task on Multilingual Lexical Simplification. In Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.
- North, K., Ranasinghe, T., Shardlow, M. *et al.* Deep learning approaches to lexical simplification: A survey. *J Intell Inf Syst* 63, 111–134 (2025).
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.